Making Connections

October 2017

# The Texas Teacher Evaluation and Support System rubric: Properties and association with school characteristics

**Valeriy Lazarev**
**Denis Newman**
**Thanh Nguyen**
**Li Lin**
**Jenna Zacamy**
Empirical Education Inc.

## Key findings

- The Texas Teacher Evaluation and Support System (T-TESS) rubric demonstrated potential to effectively differentiate teacher performance and served its purpose of yielding meaningful feedback that can support targeted professional development. More than 25 percent of teachers were rated developing or in need of improvement. Sixty-eight percent of teachers were rated proficient. Five percent of teachers received either an accomplished or a distinguished rating.
- The T-TESS rubric is internally consistent at both the domain and dimension levels. All correlations between domain ratings and all correlations between dimension ratings are positive, suggesting that none of the domains or dimensions stands out as unrelated or contradictory to the rest of the rubric.
- The T-TESS rubric is efficient. None of the domains or dimensions is clearly redundant. And each dimension makes a unique contribution to a teacher's overall rating.
- Although relationships between teachers' overall ratings on the T-TESS rubric and some school characteristics are statistically significant, the combination of school characteristics included in the analysis explains, at most, 8 percent of the variation in overall ratings for teachers in high schools and less of the variation for teachers in elementary and middle schools.

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

REL SOUTHWEST
Regional Educational Laboratory
At SEDL

**U.S. Department of Education**
Betsy DeVos, *Secretary*

**Institute of Education Sciences**
Thomas W. Brock, *Commissioner for Education Research*
*Delegated the Duties of Director*

**National Center for Education Evaluation and Regional Assistance**
Ricky Takai, *Acting Commissioner*
Elizabeth Eisner, *Acting Associate Commissioner*
Amy Johnson, *Action Editor*
Chris Boccanfuso, *Project Officer*

REL 2018–274

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 2017

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0012 by Regional Educational Laboratory Southwest administered by SEDL. 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., Nguyen, T., Lin, L., & Zacamy, J. (2017). *The Texas Teacher Evaluation and Support System rubric: Properties and association with school characteristics* (REL 2018–274). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved fromp http://ies.ed.gov/ncee/edlabs.

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

## Summary

Federal initiatives and other research have led states across the nation to sharpen their focus on teacher evaluation in recent years. In 2009 a seminal report, *The Widget Effect*, from The New Teacher Project revealed that in districts using a binary rating system to evaluate teachers, less than 1 percent of teachers received an unsatisfactory rating (Weisberg, Sexton, Mulhern, & Keeling, 2009). The remaining 99 percent were, in effect, like widgets, undifferentiated as individual professionals. Since then, a growing body of research on teacher evaluation systems has indicated that classroom observation ratings often cluster around the middle or high end of evaluation scales (Kraft & Gilmour, 2016; Lazarev & Newman, 2015). The research has also found that observation ratings are susceptible to several biases, such as incoming student achievement and school, classroom, and rater characteristics (Chaplin, Gill, Thompkins, & Miller, 2014; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014). Thus many education decisionmakers lack sufficient information to make personnel decisions and to effectively support teacher growth and development (Lazarev, Newman, & Sharp, 2014; Lipscomb, Chiang, & Gill, 2012; Mihaly & McCaffrey, 2014; Weisberg et al., 2009).

Texas is among the states that have identified teacher evaluation and support as a high priority for education policy. In 2014/15 the Texas Education Agency piloted the Texas Teacher Evaluation and Support System (T-TESS) in 57 school districts. The pilot was followed by a refinement phase in 2015/16 and statewide rollout in 2016/17. During the pilot year teacher overall ratings were based solely on rubric ratings on 16 dimensions across four domains (planning, instruction, learning environment, and professional practices and responsibilities), although T-TESS also includes a student growth measure that was piloted simultaneously.

The study examined the statistical properties of the T-TESS rubric from the 2014/15 pilot year to explore the extent to which it differentiates teachers on teaching quality and to investigate its internal consistency and efficiency. The study also explored the relationships between rubric ratings and school characteristics to investigate whether certain types of schools have teachers with higher or lower ratings.

Among the key findings:

- The T-TESS rubric demonstrated potential to effectively differentiate teacher performance and served its purpose of yielding meaningful feedback that can support targeted professional development. More than 25 percent of teachers were rated developing or in need of improvement. Sixty-eight percent of teachers were rated proficient. Five percent of teachers received either an accomplished or a distinguished rating.
- The T-TESS rubric is internally consistent at both the domain and dimension levels. All correlations between domain ratings and all correlations between dimension ratings are positive, suggesting that none of the domains or dimensions stands out as unrelated or contradictory to the rest of the rubric.
- The T-TESS rubric is efficient. None of the domains or dimensions is clearly redundant. And each dimension makes a unique contribution to a teacher's overall rating.
- Although relationships between teachers' overall ratings on the T-TESS rubric and some school characteristics are statistically significant, the combination of

school characteristics included in the analysis explains, at most, 8 percent of the variation in overall ratings for teachers in high schools and less of the variation for teachers in elementary and middle schools.

The study's findings have several implications for practice and research. The findings suggest that the T-TESS rubric demonstrates potential to be an effective, consistent, and efficient evaluation rubric. Thus, the Texas Education Association and local education agencies have a promising tool for providing evidence-based feedback and targeted professional development. Future research could try to validate ratings based on the T-TESS rubric against other measures of teacher effectiveness (for example, student growth). Such validation could shed light on whether a relationship exists between rubric ratings and a teacher's contribution to student achievement. Moreover, future studies could explore whether relationships exist between the T-TESS rubric and classroom and district characteristics. Such analysis may unearth the extent to which effective teachers are equally distributed within schools and within and across districts. Finally, further research could explore whether implementing teacher evaluation systems translates into improvements in teacher effectiveness or in long-term outcomes, such as teacher retention and student achievement.

# Contents

## Tables

## Why this study?

Over the past several years a growing body of research on teacher evaluation systems, coupled with new federal initiatives, has catalyzed the reform of teacher evaluation in many states across the nation. In 2009 *The Widget Effect,* a report from The New Teacher Project, highlighted a critical issue that has long faced the nation's schools: the inability to differentiate teachers by their performance. The report showed that in districts that used a binary rating system to evaluate teachers, less than 1 percent of teachers received an unsatisfactory rating. Other studies of teacher evaluation systems indicate that classroom observation ratings often cluster around the middle or high end of an evaluation scale (Kraft & Gilmour, 2016; Lash, Tran, & Huang, 2016; Lazarev & Newman, 2015; Weisberg et al., 2009) and that observation ratings are susceptible to several biases, such as prior student achievement and school, classroom, and rater characteristics (Chaplin et al., 2014; Steinberg & Garrett, 2016; Whitehurst et al., 2014). (See appendix A for a literature review.) Thus many education decisionmakers lack sufficient information to make personnel decisions and to effectively support teacher growth and development (Lazarev et al., 2014; Lipscomb et al., 2012; Mihaly & McCaffrey, 2014; Weisberg et al., 2009).

***During a pilot implementation in 2014/15 the Texas Teacher Evaluation and Support System (T-TESS) was used to gather information on teacher effectiveness through multiple performance metrics based on an evaluation rubric with 16 dimensions organized into four domains***

Spurred by competition for Race to the Top program funds and by the pursuit of a waiver from the No Child Left Behind Act provisions for school accountability[1] (Doherty & Jacobs, 2013), many states and districts have invested in comprehensive teacher evaluation systems that include multiple measures of teacher effectiveness, such as classroom observations, student growth measures, and teacher self-assessments. State and district policies on teacher evaluation continue to evolve following the passage of the Every Student Succeeds Act in 2015. As states and districts navigate this landscape and deliberate on the course of action for evaluating educators, generating evidence to help inform those decisions remains a high priority.

Texas is among the states that have identified teacher evaluation and support as a high priority for education policy. Beginning in fall 2013 the Texas Education Agency began to develop the Texas Teacher Evaluation and Support System (T-TESS) in order to shift away from the previous system, the Professional Development and Appraisal System, which had been in place since 1997 and was used in 86 percent of the state's districts. Under the Professional Development and Appraisal System less than 2 percent of teachers received a below proficient rating (Ettema, Sengupta, & Kress, 2014). That outcome fueled the need for the new system, which aimed to "improve […] the quality of individual teacher evaluations so that teachers and administrators have more meaningful feedback on student learning and growth" (Texas Education Agency, 2016a, p. 3). During a pilot implementation in 2014/15 T-TESS was used to gather information on teacher effectiveness through multiple performance metrics based on an evaluation rubric with 16 dimensions organized into four domains (see table B1 in appendix B), and a student growth measure.

As T-TESS was being developed, the Regional Educational Laboratory Southwest Educator Effectiveness Research Alliance[2] and the Texas Education Agency requested a study that would help the agency identify potential improvements for scoring guidelines, rescaling items, changing weightings, or excluding items. In Texas the state can recommend an evaluation system, but districts retain the option to select their own. T-TESS has replaced the Professional Development and Appraisal System as the state-recommended evaluation system (Texas Administrative Code, Chapter 150, Section 150.1001). So in addition

to supporting the Texas Education Agency, the findings and methodology of the current study may also be of interest to other state and local education agencies, as well as policymakers and educators in Texas and beyond who are developing or implementing multiple-measure teacher evaluation systems.

## What the study examined

In 2014/15 T-TESS was piloted in 57 districts across the state, and in 2015/16 the pilot expanded to 200 districts as part of a refinement phase in preparation for statewide rollout in 2016/17. The current study analyzed ratings on the T-TESS rubric (rubric ratings) for 8,250 teachers across 251 schools and 51 districts[3] from the pilot. The study aimed to garner insight into the statistical properties of the rubric and to investigate potential biases from school characteristics. During the pilot teacher overall ratings were based solely on the rubric ratings, although T-TESS also included a student growth measure that was piloted simultaneously.

***The current study was intended to help the Texas Education Agency identify potential improvements for scoring guidelines, rescaling items, changing weightings, or excluding items***

The study addressed four research questions that aimed to provide the Texas Education Agency with information on how well the T-TESS rubric measures teacher effectiveness, which may be relevant and useful for future refinement of the rubric:

- To what extent do overall, domain, and dimension ratings on the T-TESS rubric differentiate teacher effectiveness?
- To what extent is the T-TESS rubric internally consistent?
- To what extent is the T-TESS rubric efficient?
- To what extent are overall and domain ratings on the T-TESS rubric associated with school characteristics?

Box 1 summarizes the data sources, sample, and methods, and appendix C provides further details. Box 2 defines key terms used in the report.

### Box 1. Data and methods

#### Data

The dataset comprised ordinal overall, domain, and dimension ratings on the rubric from the Texas Teacher Evaluation and Support System (T-TESS), referred to here as rubric ratings, for 8,250 teachers across 251 schools and 51 districts from the 2014/15 pilot as well as data on school characteristics.

For the data used in this study for each teacher, the Texas Education Agency converted each of the 16 dimension ratings into a score on a five-point numerical scale (from 1 = improvement needed to 5 = distinguished) and then averaged the scores for the dimensions within each of the four domains to generate four domain scores. Overall scores were the average of the domain scores.[1] Overall and domain scores were converted to an ordinal rating: scores below 2.0 became a rating of improvement needed, scores of 2.0–2.99 became a rating of developing, scores of 3.0–3.99 became a rating of proficient, scores of 4.0–4.99 became a rating of accomplished, and scores of 5.0 became a rating of distinguished.

Data on school characteristics were from the Texas Education Agency's Texas Academic Performance Report database for the 2014/15 school year. The characteristics included in the analysis fell into four categories:

- General profile information: grade span, school locale, and number of students.
- Demographic information: racial/ethnic distribution of students, percentage of students eligible for the federal school lunch program, percentage of students who are English learner students, and percentage of students in special education.

*(continued)*

## Box 1. Data and methods *(continued)*

- Achievement information: percentage of students who receive at least a satisfactory rating on the State of Texas Assessments of Academic Readiness reading test for students in grades 3–8 and school academic distinctions received (see box 2).
- Teacher information: percentage of teachers with five or fewer years of experience and percentage of teachers with a master's or doctoral degree.

School-level characteristics were used instead of classroom-level characteristics because ratings were deidentified at the teacher level and therefore could not be linked to classroom-level data.

### Methods

To explore the extent to which rubric ratings differentiate teacher effectiveness (research question 1), the study team calculated descriptive statistics, including the distribution of overall, domain, and dimension ratings. To examine the rubric's internal consistency (research question 2), the study team calculated pairwise correlations between the four domain scores and between the 16 dimension ratings. To assess the rubric's efficiency (research question 3), the study team reused the correlation results from research question 2 and examined uniqueness values. To examine the relationship between rubric ratings and school characteristics (research question 4), the study team conducted group comparisons and performed five sets of linear regression analyses with teacher-level rubric ratings as outcome variables and school characteristics as covariates, with school and district random effects to account for correlated standard errors. The first set of models used the overall ratings as the outcome variable, and the four subsequent sets used the four domain ratings. For each set the first model included school characteristics that the study team hypothesized, on the basis of past research, could relate to teacher evaluation ratings (see appendix A for a literature review). The second model was a reduced model, which was less complex (had fewer covariates) than the full model. The third, fourth, and fifth models employed the subsample of elementary, middle, and high schools, respectively.

### Note

**1.** About 2 percent of the overall and domain ratings were not simple averages of the underlying ratings. See appendix C for details. Districts were not required to convert ordinal ratings into numeric values when providing feedback to teachers.

## Box 2. Key terms

***Academic distinction designation.*** Recognition awarded to schools that are among the top 25 percent on various performance indicators compared with 40 similar schools. Performance indicators depend on the grade level and include attendance rates, performance on state assessments, and student participation in the ACT, Advanced Placement courses, International Baccalaureate courses, and the SAT (Texas Education Agency, 2015).

***Advantaged school.*** A school that is in the bottom quintile of percentage of students eligible for the federal school lunch program and that received an academic distinction designation in English language arts/reading and math.

***Disadvantaged school.*** A school that is in the top quintile of percentage of students eligible for the federal school lunch program and that did not receive an academic distinction designation.

***Efficiency.*** The degree to which an instrument is composed of items that contribute unique information.

***Internal consistency.*** A measure of the extent to which items (dimensions or domains in the current study) that propose to measure the same construct correlate with each other.

***Uniqueness.*** A statistical metric produced by factor analysis that represents the proportion of variance of a variable (dimension in the current study) that cannot be attributed to any other variables in the model (that is, the proportion of variance that is unique to the variable). The uniqueness value ranges from 0 (fully correlated with other dimensions already measured and therefore redundant) to 1 (not at all correlated with other dimensions; Cattell, 1973; Kim & Mueller, 1978).

## What the study found

This section presents the key findings of the study.

### The rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot differentiates teacher effectiveness at the overall, domain, and dimension levels

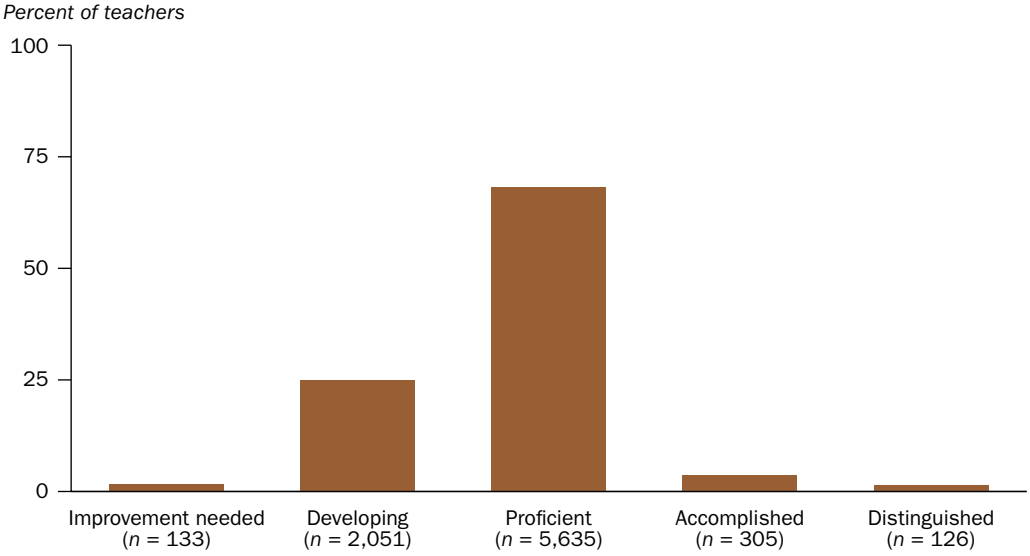
The T-TESS rubric demonstrated potential in effectively differentiating teacher performance and served its purpose of yielding meaningful feedback that can support targeted professional development. More than 25 percent of teachers were rated developing (24.9 percent) or in need of improvement (1.6 percent; figure 1). Five percent of teachers rose above the norm and received either an accomplished (3.7 percent) or a distinguished rating (1.5 percent). The remaining teachers (68.3 percent) received a proficient rating as an overall rating.

*The T-TESS rubric demonstrated potential in effectively differentiating teacher performance and served its purpose of yielding meaningful feedback that can support targeted professional development*

The distribution of ratings differed slightly across domains (figure 2). The ratings distributions in the planning and instruction domains were most similar, although the instruction domain had a higher percentage of teachers who received a developing rating. The learning environment domain had a lower percentage of teachers who received a proficient rating but a higher percentage who received an accomplished or distinguished rating. The professional practices and responsibilities domain had the highest percentage of teachers who received a proficient rating.

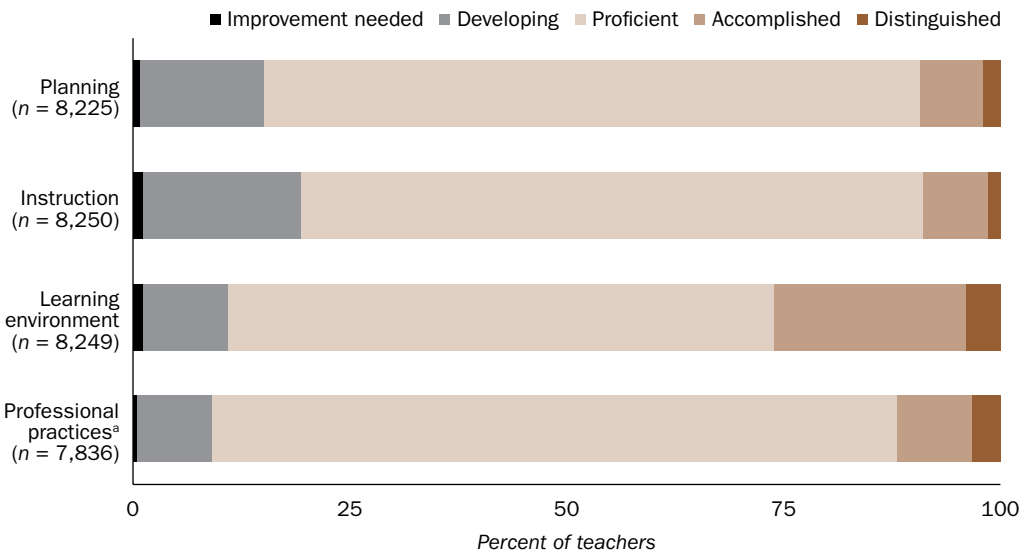
The learning environment domain exhibited the widest distribution of ratings (0.70 standard deviation), meaning that ratings for that domain varied the most across teachers (table 1). The distribution was narrower for the instruction (0.59 standard deviation), planning (0.57), and professional practices and responsibilities (0.56) domains, meaning that within a domain teachers were more likely to receive a rating that is close to the mean.

**Figure 1. The Texas Teacher Evaluation and Support System rubric from the 2014/15 pilot demonstrated potential in effectively differentiating teacher performance**




**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Figure 2. The distribution of rubric ratings from the 2014/15 Texas Teacher Evaluation and Support System pilot differed slightly across domains**




***Ratings for the learning environment domain varied the most across teachers***

**a.** Two districts did not provide data for the professional practices and responsibilities domain (see appendix C).

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

The learning environment domain had the highest mean score (3.2), followed closely by the professional practices and responsibilities domain (3.1), the planning domain (3.0), and the instruction domain (2.9; see table 1).

The mean of dimension scores ranged from 3.1 points to 3.4 points (see table 1). The standard deviation of dimension scores ranged from 0.53 (for the professional development dimension within the professional practices and responsibilities domain) to 0.72 (for the managing student behavior dimension within the learning environment domain). In other words, teachers were more likely to receive the same rating for the professional development dimension than they were for the managing student behavior dimension. This is also suggested by the fact that 77 percent of teachers received a proficient rating for the professional development dimension, compared with 50 percent for the managing student behavior dimension.

### The rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot is internally consistent

Correlations between domain ratings on the rubric from the 2014/15 T-TESS pilot were all positive and statistically significant, ranging from .47 (between the professional practices and responsibilities domain and the learning environment domain) to .72 (between the instruction domain and the planning domain; table 2). That the professional practices and responsibilities domain had the lowest correlations with the other domains was expected because it was rated separately from the other three domains at an end-of-year conference with the teacher rather than during classroom observation. That correlations were highest between the planning domain and the instruction domain supports the later finding that the two domains measure a common, underlying element of teacher effectiveness.

## Table 1. Descriptive statistics for rubric ratings from the 2014/15 Texas Teacher Evaluation and Support System pilot, by domain and dimension

| Domain and dimension | Number of teachers | Mean score[a] | Standard deviation | Percent receiving a proficient rating |
|---|---|---|---|---|
| Domain 1: Planning | 8,225 | 3.0 | 0.57 | 76 |
| 1.1 Standards and alignment | 8,219 | 3.2 | 0.59 | 68 |
| 1.2 Data and assessment | 8,198 | 3.1 | 0.59 | 72 |
| 1.3 Knowledge of students | 8,211 | 3.3 | 0.63 | 63 |
| 1.4 Activities | 8,195 | 3.2 | 0.67 | 60 |
| Domain 2: Instruction | 8,250 | 2.9 | 0.59 | 72 |
| 2.1 Achieving expectations | 8,247 | 3.2 | 0.63 | 66 |
| 2.2 Content knowledge and expertise | 8,246 | 3.3 | 0.68 | 56 |
| 2.3 Communication | 8,233 | 3.2 | 0.64 | 62 |
| 2.4 Differentiation | 8,248 | 3.1 | 0.69 | 62 |
| 2.5 Monitor and adjust | 8,213 | 3.2 | 0.65 | 64 |
| Domain 3: Learning environment | 8,249 | 3.2 | 0.70 | 63 |
| 3.1 Classroom environment, routines, and procedures | 8,246 | 3.4 | 0.70 | 50 |
| 3.2 Managing student behavior | 8,243 | 3.4 | 0.72 | 50 |
| 3.3 Classroom culture | 8,234 | 3.4 | 0.70 | 54 |
| Domain 4: Professional practices and responsibilities[b] | 7,836 | 3.1 | 0.56 | 79 |
| 4.1 Professional demeanor and ethics[b] | 7,834 | 3.4 | 0.71 | 60 |
| 4.2 Goal setting[b] | 7,823 | 3.2 | 0.55 | 73 |
| 4.3 Professional development[b] | 7,823 | 3.2 | 0.53 | 77 |
| 4.4 School community involvement[b] | 7,804 | 3.3 | 0.63 | 67 |

**Note:** The rubric from the Texas Teacher Evaluation and Support System pilot uses a five-point scale for rubric ratings: 1 = improvement needed, 2 = developing, 3 = proficient, 4 = accomplished, and 5 = distinguished. At each rubric rating level the minimum score was 1, the maximum score was 5, and the mode was 3.

**a.** Domain mean scores are lower than their dimension mean scores because districts had the option to convert ordinal values into numerical values and to average them as described in box 1 (see also appendix C).

**b.** Two districts did not provide data for domain 4 (see appendix C).

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

## Table 2. Correlations between domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot

| Domain | 1. Planning | 2. Instruction | 3. Learning environment |
|---|---|---|---|
| 1. Planning | | | |
| 2. Instruction | .72 | | |
| 3. Learning environment | .57 | .62 | |
| 4. Professional practices and responsibilities | .53 | .50 | .47 |

**Note:** All correlation coefficients are statistically significant at $p < .001$.

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

The correlations between dimension ratings within each domain were also all positive and statistically significant (see table E1 in appendix E). The correlations between dimension ratings within a domain were higher for the learning environment domain (.62–.65) than for the other three domains (.47–.58), suggesting a greater potential for redundancies within that domain. This supports the later finding that dimensions within the learning environment domain had the lowest uniqueness values (see table E2 in appendix E). The lowest correlation (.47) was between the data and assessment dimension and the activities dimension within the planning domain, but the range of correlations between dimensions within that domain (.47–.53) was similar to that between dimensions within the instruction domain (.48–.58) and that between dimensions within the professional practices and responsibilities domain (.49–.55).

*The correlations between domain ratings and between dimension ratings within each domain were all positive and statistically significant*

### The rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot is efficient, with each dimension making a unique contribution to a teacher's overall rating

That none of the correlations between domain ratings and none of the correlations between dimensions ratings were close to 1 (see table 2 and table E1 in appendix E) suggests that no two dimensions and no two domains on the rubric from the 2014/15 T-TESS pilot were clearly redundant.

Exploratory factor analysis shows that uniqueness values of dimensions are within a narrow range of .33–.53 for a two-, three-, or four-factor model (see table E2 in appendix E). No dimension stood out as a distinctively low- or high-uniqueness item. There is thus no clear indication that any of the dimensions could have been redundant. In other words, most dimensions made a unique contribution to a teacher's overall rating.

### Although relationships between overall ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot and some school characteristics are statistically significant, the combination of school characteristics included in the analysis explains, at most, 8 percent of the variation in overall ratings

There were small (0.05–0.13 point) but statistically significant differences in the average overall ratings on the rubric from the 2014/15 T-TESS pilot across school characteristics (see table E3 in appendix E). The difference was most pronounced between teachers in schools in the bottom quintile of percentage of students eligible for the federal school lunch program and teachers in schools in the top quintile (0.13 point) and between teachers in schools in the bottom quintile of percentage of students in special education and teachers in schools in the top quintile (0.11 point). In other words, teachers in schools with a lower percentage of students eligible for the federal school lunch program and schools with a lower percentage of students in special education tended to receive higher rubric ratings. Similarly, teachers in advantaged schools (as defined in box 2) and schools in the bottom quintile of percentage of teachers with five or fewer years of experience tended to receive higher rubric ratings (see tables E3 and E4 in appendix E).

At the domain level teachers in schools in the bottom quintile of percentage of students eligible for the federal school lunch program received a rating that was 0.12–0.16 point higher, on average, than teachers in schools in the top quintile (see table E5 in appendix E).[4] The pattern was similar for other school characteristics: percentage of students who are racial/ethnic minority students, percentage of students who are English learner

students, and percentage of students in special education. Teachers in schools in the bottom quintiles of percentage of racial/ethnic minority students, percentage of English learner students, and percentage of students in special education and teachers in advantaged schools tended to receive significantly higher rubric ratings (see table E5 and E6 in appendix E) than teachers in schools in the top quintiles and teachers in disadvantaged schools. For the learning environment domain and the professional practices and responsibilities domain, rubric ratings did not differ between teachers in schools in the top quintile percentage of teachers with five or fewer years of experience and teachers in schools in the bottom quintile percentage.

The results of the linear regression provided a more refined view into the results of the subgroup analysis. Model 1 (the full model) showed that after other school characteristics are controlled for, schools with a higher percentage of students eligible for the federal school lunch program had a statistically significant and negative association with overall teacher rubric ratings (table 3). On average, a 1 percentage point increase in the percentage of eligible students was associated with a 0.006 point decrease in a teacher's rubric score.[5] Conversely, schools with more students and schools with a higher percentage of English learner students had a positive association with teacher rubric ratings. A 1 percentage point increase in the percentage of English learner students was associated with a 0.004 point increase in a teacher's rubric score. However, statistically significant differences may not be substantively important differences.[6]

***On models disaggregated by school grade span, the relationships between overall rubric ratings and school characteristics differed by school grade span***

Overall rubric ratings were 0.117 point higher, on average, for teachers in elementary schools than for teachers in high schools, though the estimate is not statistically significant (see table 3). There was no significant difference in overall rubric ratings between teachers in urban schools and teachers in town, suburban, or rural schools.

Model 2 (the reduced model, which omitted school grade span, school locale, percentage of students who are White, percentage of students in special education, percentage of teachers with five or fewer years of experience, and percentage of teachers with a master's or doctoral degree as covariates) explained the same percentage of the variation in rubric ratings as model 1 (approximately 4 percent; see table 3).

Models 3–5 (which used the full model on samples disaggregated by school grade span—elementary, middle, and high school) showed that the relationships between overall rubric ratings and school characteristics differed by school grade span (see table 3). For teachers in elementary schools overall ratings were positively associated with the percentage of English learner students (0.005 point) and the percentage of teachers with a master's or doctoral degree (0.007 point). For teachers in middle schools overall ratings were positively associated with the percentage of students who received at least a satisfactory rating on the State of Texas Assessments of Academic Readiness reading test (0.024 point). And for teachers in high schools overall ratings were positively associated with school size (0.206 point for an increase of 1,000 students) and negatively associated with the percentage of students eligible for the federal school lunch program (–0.009 point). Overall ratings were not associated with school locale, percentage of racial/ethnic minority students, percentage of students in special education, and percentage of teachers with five or fewer years of experience for teachers in elementary, middle, or high school.

**Table 3. Regression results for the relationship between overall ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot and school characteristics**

| Intercept and covariate | Model 1: full sample, full model | Model 2: full sample, reduced model | Model 3: elementary school sample, full model | Model 4: middle school sample, full model | Model 5: high school sample, full model |
|---|---|---|---|---|---|
| Intercept | 3.036*** (0.333) | 2.731*** (0.23) | 2.860*** (0.515) | 1.201 (0.925) | 3.216*** (0.586) |
| School grade span (high school is reference group) | | | | | |
| Elementary | 0.117 (0.074) | | | | |
| Middle | 0.058 (0.073) | | | | |
| School locale (urban is reference group) | | | | | |
| Suburb | –0.034 (0.084) | | 0.032 (0.11) | –0.316. (0.158) | 0.111 (0.153) |
| Town | –0.032 (0.094) | | 0.007 (0.124) | –0.284 (0.191) | 0.288 (0.186) |
| Rural | –0.016 (0.093) | | 0.003 (0.124) | –0.301. (0.175) | 0.303 (0.185) |
| School size | | | | | |
| Number of students (unit of change: 1,000 students) | 0.125* (0.049) | 0.112** (0.042) | 0.167 (0.165) | 0.134 (0.191) | 0.206** (0.061) |
| School demographics (unit of change: 1 percentage point) | | | | | |
| Percentage of students who are White | 0.001 (0.001) | | 0.002 (0.002) | –0.004 (0.003) | 0.001 (0.003) |
| Percentage of students eligible for the federal school lunch program | –0.006** (0.002) | –0.005*** (0.001) | –0.005 (0.003) | –0.000 (0.005) | –0.009* (0.005) |
| Percentage of English learner students | 0.004* (0.002) | 0.004** (0.001) | 0.005* (0.002) | 0.003 (0.005) | 0.008 (0.006) |
| Percentage of students in special education | –0.002 (0.008) | | 0.000 (0.011) | –0.001 (0.017) | 0.004 (0.015) |
| School achievement | | | | | |
| Percentage of students scoring proficient on the State of Texas Assessments of Academic Readiness reading test (unit of change: 1 percentage point) | –0.001 (0.003) | 0.003 (0.002) | –0.001 (0.004) | 0.024** (0.008) | –0.006 (0.005) |
| School teacher information (unit of change: 1 percentage point) | | | | | |
| Percentage of teachers with five or fewer years of experience | –0.002 (0.002) | | –0.002 (0.002) | 0.002 (0.003) | –0.000 (0.004) |
| Percentage of teachers with a master's or doctoral degree | 0.002 (0.002) | | 0.007* (0.003) | –0.006 (0.004) | –0.002 (0.004) |
| | | | | | |
| Number of teachers | 7,387 | 7,387 | 3,404 | 1,764 | 2,219 |
| Number of schools | 222 | 222 | 120 | 53 | 49 |
| Number of districts | 50 | 50 | 43 | 35 | 34 |
| Akaike information criterion | 12,310 | 12,256 | 5,491 | 3,066 | 3,846 |
| Adjusted *R*-squared | 0.04 | 0.04 | 0.05 | 0.06 | 0.08 |

* Significant at p < .05; ** significant at p < .01; *** significant at p < .001.

**Note:** Numbers in parentheses are standard errors.

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

The variation in overall rubric ratings explained by the combination of school characteristics (adjusted R-squared) differed slightly across models for overall ratings and for each domain. For overall ratings the highest adjusted R-squared was for model 5, meaning that the combination of school characteristics included in the analyses best predicted ratings for the high school sample, explaining about 8 percent of the variation, compared with 6 percent for the middle school sample and 5 percent for the elementary school sample. The pattern across school grade spans was similar for each domain, except for the professional practices and responsibilities domain, where the combination of school characteristics best predicted ratings for middle schools (9 percent), followed by elementary schools (8 percent) and high schools (3 percent; see tables E7–E10 in appendix E).

Because this was a correlational analysis, the findings indicate that a relationship between rubric ratings and certain school characteristics exists but do not imply causality in either direction. Whether the differences in ratings across grade spans were due to actual differences in teacher effectiveness or to other reasons, such as the instrument, the rater, or the type of students, cannot be determined.

## Implications of the study findings

The study findings have several promising implications for practice and future research.

***The distribution of ratings may help inform teacher support strategies***

The study found that the rubric differentiated between teachers with a proficient rating (68 percent) and a developing rating (25 percent), that about 5 percent of teachers received an accomplished or distinguished rating, and that 1.6 percent received the lowest rating of improvement needed. The Texas Education Agency and other local and state stakeholders may want to compare the results of this study to their own expectations of the distribution of teacher effectiveness based on field experience. Any gap could be explored to identify the source of the discrepancies, such as issues related to the implementation of the rubric or to cultural forces that compel appraisers to provide certain ratings (Weisberg et al., 2009). The distribution of ratings may also help inform teacher support strategies. For example, because teachers tend to rate lower in both the planning and instruction domains, decisionmakers could investigate whether teachers would benefit from greater support in those areas.

The study found that the professional practices and responsibilities domain had the lowest correlation with the other three domains. This could be due partly to the fact that the professional practices and responsibilities domain, unlike the other three domains, is observed outside the classroom. However, this finding could benefit from deeper exploration into whether teacher effectiveness related to planning, instruction, and the learning environment domains actually differs from teacher effectiveness related to the professional practices and responsibilities domain or whether the low correlation can be attributed to the rubric itself or to how raters are trained to code dimensions and domains.

The study found that the rubric is efficient, meaning that each dimension makes a unique contribution to a teacher's overall rating. Because the finding suggests that the rubric does not need to be further reduced, future research could examine the practicality and feasibility of continuing to administer the rubric in its entirety and the burden that doing so places on the rater.

While the study found that the combination of school characteristics included in the analysis can explain, at most, 8 percent of the variation in overall rubric ratings, the results suggest some differences in how the rubric functions for teachers in different grade spans. Because differences in ratings may not be due to differences in teaching quality, potential biases can be minimized by making raters aware of the nuances and complexities of teaching at each grade span, by statistically adjusting the ratings after the fact, or by adjusting the rubric so it functions the same way across grade spans.

The study's findings can also serve as departure points for future research on teacher effectiveness. One direction is to validate the rubric ratings against other measures of teacher effectiveness that are closer to student outcomes, including student growth measures. Such validation could shed light on how ratings are related to teachers' contributions to student improvement, at least in grades and subjects where standardized measures are available. It could also illuminate the many possible explanations for these findings, which may be attributed to the instrument, the rater, or real differences in teaching quality (Cohen & Goldhaber, 2016).

***The study found that the rubric is efficient, meaning that each dimension makes a unique contribution to a teacher's overall rating***

Future research could also broaden the investigation to whether rubric ratings relate to classroom and district characteristics in similar magnitude and direction as they do to school characteristics.[7] Such analysis could shed light on the dynamics of equitable distribution of measured teacher effectiveness within schools and within or across districts, as observed in the literature on teacher sorting (Goldhaber, Lavery, & Theobald, 2015; Kalogrides, Loeb, & Béteille, 2013). Because of data limitations, this study focuses on the relationship between rubric ratings and school characteristics for only one year, and the findings may have been affected by factors specific to that year. Data for multiple years might allow researchers to single out persistent, idiosyncratic school effects to improve the accuracy of estimates.

Finally, future research could benefit from a deeper understanding of how implementation of the system affects short- and long-term outcomes, such as teacher retention, teacher growth, and improved student achievement.

## Limitations of the study

The study has four main limitations.

First, the data for the study were collected during a stage in the T-TESS pilot when both implementation and data collection processes were still emerging and evolving. The processes were likely less standardized and the documentation less readily available than they would have been for a system that had been implemented over a longer period. However, on the basis of communication with the Texas Education Agency and the National Institute of Education and Training about uncertainties regarding the data of rubric ratings, the study team believes that it collected enough information about the pilot context in which to ground the study's analysis and results.

Second, the sample included about 5 percent of Texas school districts and 3 percent of Texas schools that volunteered to participate over one year (2014/15). Although this sample size limits the generalizability of the study findings, it contains enough districts and schools to obtain reasonably precise estimates. But even reasonably precise estimates

may not predict outcomes in the statewide rollout because participation in the pilot was voluntary and there may have been differential attrition and subject noncompliance (that is, district or school administrators and teachers may not have completed the entire evaluation protocol). Suburban and town districts are better represented in the pilot sample than urban and rural districts, and the pilot sample averages more teachers per school than the state as a whole (see appendix D), which limits the generalizability of the findings.

Third, the majority of teachers were observed only once. Ideally, each teacher would have been rated several times by several raters. This would have enabled the study team to calculate interrater reliability and test–retest reliability. However, the nature of the pilot constrained the amount of time that participants could devote to data collection, thus limiting the scope of the study and the precision of estimates.

Fourth, because the rubric ratings were deidentified at the teacher level, relationships between rubric ratings and characteristics could be investigated only at the school level rather than at the classroom level. Given the heterogeneity across classrooms within schools, the aggregation of data to the school level may have resulted in the loss of some information.

***Although the sample size limits the generalizability of the study findings, the sample contains enough districts and schools to obtain reasonably precise estimates***

## Appendix A. Literature review

States and school districts across the country are overhauling their teacher evaluation systems, shifting the focus from teacher qualifications to teacher effectiveness. Until recently, states have relied on academic credentials and years of experience to make personnel decisions (Whitehurst et al., 2014). However, more states are moving toward comprehensive, multifaceted educator evaluation policies, spurred by competition for Race to the Top program funds and by the pursuit of the No Child Left Behind Act of 2001 waivers (Doherty & Jacobs, 2013). State and district policies around teacher evaluations continue to evolve with the recent passage of the Every Student Succeeds Act, a reauthorization of the Elementary and Secondary Education Act. As states and districts navigate this landscape and deliberate on the course of action for evaluating educators, it is important to continue to generate evidence to help inform decisions.

A major catalyst for this shift in national priorities was the release of *The Widget Effect*, a 2009 report by The New Teacher Project (Weisberg et al., 2009). Conducted in 12 districts across four states, the study called attention to a national crisis—the inability of schools to effectively differentiate between low- and high-performing teachers. In districts that used binary evaluation ratings, more than 99 percent of teachers received a satisfactory rating. In districts that used a broader range of rating options, less than 1 percent of teachers received an unsatisfactory rating. Teachers were, in effect, like widgets, undifferentiated as individual professionals. Consequently, excellent teachers went unacknowledged, poor performance was not addressed, and professional development, especially for new, inexperienced teachers, was not targeted toward areas of need and was thus inadequate. The report challenged states to adopt a comprehensive performance evaluation system that can inform critical personnel decisions, including teacher assignment, professional development, compensation, retention, and dismissal (Weisberg et al., 2009).

Despite recent efforts to implement new teacher evaluation and support programs, questions remain about how to assign weights to the various components of evaluation systems, the validity of component metrics, and optimal approaches for calculating summative rankings. The literature on teacher evaluation systems is rich with policy briefs and guides on how to design teacher evaluation systems (Darling-Hammond, 2015; Doherty & Jacobs, 2013; Hull, 2013; The New Teacher Project, 2010). This appendix focuses on seminal projects and research studies that are most relevant to the research questions of the current study.

The following sections discuss differentiation of teacher effectiveness, internal consistency and efficiency of teacher evaluation instruments, and the relationships between evaluation ratings and school characteristics. Classroom observation ratings receive special emphasis because they play a significant role in teacher evaluation systems. Because the No Child Left Behind Act requires tests only in math and English language arts, many teachers —80 percent, according to Whitehurst et al. (2014)—do not have students who take state tests, and thus classroom observations are the main contribution to their evaluation score.

### Differentiation of teacher effectiveness

A key message of *The Widget Effect* was the need to differentiate between teachers on a continuum of instructional quality (Weisberg et al., 2009). Between 2009 and 2013 the

number of states that changed their requirements from having two categories of evaluation ratings (satisfactory and not satisfactory) to more than two increased from 17 to 43. Although using multiple category ratings does not necessarily translate into greater differentiation of teacher effectiveness, it does lay important groundwork for consideration (Doherty & Jacobs, 2013). This section discusses the current evidence of the ability of states and school districts to differentiate teacher effectiveness.

Lazarev et al. (2014) examined the statistical properties of a pilot teacher evaluation system in five Arizona school districts in 2012/13. The data included item-level results[8] for teacher observations from two observation cycles of Charlotte Danielson's Framework for Teaching (FFT); student academic progress calculations from state and supplementary tests; and summative results from student, parent, and peer surveys. The study found that ratings on observation items and components shared similar features—their distribution was heavily concentrated around the median and skewed toward higher ratings. Most teachers received a proficient rating on most observation items; only 2 percent received an unsatisfactory rating. The authors concluded that such concentration of ratings provides insufficient information for decisionmakers to distinguish between the highest and lowest performers for purposes of professional development or administrative decisions, assuming that teaching quality does in fact vary in the study sample.

Chaplin et al. (2014) examined a new teacher evaluation system in Pittsburgh Public Schools during the 2011/12 school year. The Pittsburgh teacher evaluation model was based on several measures: the Research-based Inclusive System of Evaluation (RISE) observation protocol, which is drawn from the FFT and relies on principals' assessment, the Tripod Student Perception Survey, and a value-added measure of student test scores for three years. For each RISE observation component a majority of teachers received a proficient rating. The authors concluded that although all three measures have the potential to differentiate among teachers, only the district's value-added measures reliably differentiate among teachers.

The findings from the Bill & Melinda Gates Foundation's (2013) Measures of Effective Teaching (MET) project—the largest study of instructional practice and its relationship to student outcomes—contrast with the results from Arizona and from Pittsburgh. The three-year project collected data from nearly 3,000 teachers across six school districts and employed several classroom observation protocols: the Classroom Assessment Scoring System (CLASS), the FFT, the Quality Science Teaching, the Protocol for Language Arts Teaching Observations (PLATO), state English language arts and math tests and supplemental tests, the Tripod Student Perception Survey, and the UTeach Teacher Observation Protocol (UTOP). Teachers were rated multiple times throughout the year by multiple external raters who were trained by Educational Testing Service.

The ratings in MET project districts were less skewed toward high ratings than the ratings in the Arizona districts, which Lazarev et al. (2014) hypothesized could be attributed to the fact that raters in the Arizona pilot received limited preparation and training and were motivated by different incentives. In Arizona, observations were usually conducted by the principal, while in the MET project they were conducted for research purposes by external raters.

Using the same observation FFT protocol as the MET project and Arizona studies, Lipscomb et al. (2012) found that principals in Pittsburgh gave 96 percent of their teachers a

proficient or distinguished rating. The results suggested that although classroom observations could differentiate teacher effectiveness, the results depended heavily on the number of raters and their characteristics, such as the rater's role relative to the teacher and whether the rater had been trained and certified (Bill & Melinda Gates Foundation, 2013). Without proper training, raters may not be cognizant of their own biases, which could influence their judgment and lead to systematic errors in scoring (Yoon, Chen, & Holtzman, 2014).

Kraft and Gilmour's (2016) compilation of teacher effectiveness ratings found that in a majority of the 19 states that have reformed their evaluation system since *The Widget Effect* was released, fewer than 3 percent of teachers received a rating below proficient. New Mexico stood out as an outlier: 26.2 percent of teachers received a rating below proficient. The authors also presented survey data from an urban district demonstrating that raters' perceptions were not reflected in the assigned ratings. Only one in three teachers whom raters perceive as below proficient actually received a rating below proficient.

### Internal consistency and efficiency

Internal consistency and efficiency are two important considerations in designing a teacher evaluation system. Internal consistency of ratings is a measure of the extent to which items that propose to measure the same construct correlate with each other. Efficiency is the degree to which an instrument is composed of items that contribute unique information. Efficiency is important because each additional competency included in an instrument adds costs. For example, adding a competency requires training and scoring time for raters. It also risks lowering the quality of data on all the other competencies because raters have already reached the limits of their ability to keep track (Kane & Staiger, 2012).

Across multiple studies correlational analysis indicated that each instrument of evaluation systems captured a common aspect of teacher effectiveness—although some indicators within and across instruments also captured distinct dimensions of teacher effectiveness (Chaplin et al., 2014; Kane & Staiger, 2012; Lazarev et al., 2014). Kane and Staiger (2012) found low correlations between the general teaching observation instruments (FFT and CLASS) and the math-specific observation instruments (Mathematical Quality of Instruction and UTOP). In contrast, they found high correlations between the general teaching observation instruments (FFT and CLASS) and the English language arts–specific instrument (PLATO). These findings suggest that the instruments were indeed measuring distinct concepts of teacher effectiveness and that the English language arts instrument measured competencies that were more aligned with the general teaching instruments than with the math instruments.

These results were supported by findings from principal component analyses (Chaplin et al., 2014; Kane & Staiger, 2012; Lazarev & Newman, 2014). For each instrument of the MET project, Kane and Staiger (2012) found that three clusters of competencies generally accounted for most of the teacher-level variation in ratings. The first principal component captured the teacher's overall performance averaged across all measures. The second depended heavily on managing classroom procedures and student behavior. The third was unique to each instrument—for example, student generation of ideas for UTOP and teachers' questioning and assessment techniques for FFT. The findings supported the claim that more than one underlying factor was driving teacher effectiveness.

Using data from 450 middle school teachers who participated in the Understanding Teacher Quality study, Lockwood, Savitsky, and McCaffrey (2015) discovered two distinct teaching constructs in math and English language arts: quality of instructional practices and quality of teacher management of classrooms. Using a condensed version of the Quality Science Teaching instrument, which measured qualities of effective science teaching practices (used in the MET project), Schultz and Pecheone (2014) found that the factors were precisely delineated by the instrument's lab-based versus non-lab-based elements.

In addition, Lazarev and Newman (2014) performed a factor analysis on 57 variables in the MET data that were collected using observation ratings, student surveys, and value-added scores. They found a three-factor model to be most appropriate. The first factor, which the authors labeled as effective, was associated with student achievement and reflects teachers' skills in following procedures and in managing classroom and student behavior. The second factor, constructive, was defined by dimensions related to pedagogical devices, such as instructional dialog, feedback, and discussion. The third factor, positive, consisted mainly of student survey items that were related to the teacher's connection to the students and students' positive feelings.

### Teacher observation ratings and student and school characteristics

A fair teacher evaluation system must award the same score to teachers who are equally effective at teaching, regardless of the context of the classroom or school. However, collecting data through classroom observations is a complex process that is susceptible to biases introduced by student, classroom, and rater characteristics. This section examines the relationships between observation ratings and student and school characteristics to explore potential biases.

Whitehurst et al. (2014) explored this relationship using 2009–12 data from four school districts. They found a positive relationship between teacher observation ratings and the classroom-average student pretest scores. The authors hypothesized that the results may be capturing observation bias. To test for bias, they adjusted teacher observation ratings by controlling for class composition demographics, including the percentage of students of different races/ethnicities, the percentage of students eligible for the federal school lunch program, and the percentage of students who have learning disabilities. The statistical adjustment reduced the association between teacher observation ratings and prior student achievement test scores. Thus, without the adjustment for student demographics, the ratings would not have been a true measure of teacher effectiveness.

Even after the adjustment, the correlation was still positive and significant. The literature on teacher sorting, which documents systematic differences in the distribution of teacher characteristics across schools serving different student populations (Kalogrides et al., 2013; Lankford, Loeb, & Wyckoff, 2002), might shed some light on this association. One hypothesis from this area of the literature is that some high-performing teachers are self-selecting into classrooms where students are high achievers (Boyd, Lankford, Loeb, & Wyckoff, 2005; Clotfelter, Ladd, & Vigdor, 2006; Hanushek, Kain, & Rivkin, 2004). Another hypothesis is that principals and other administrators are assigning certain teachers to higher performing classes (Clotfelter, Ladd, & Vidgor, 2005; Feng, 2010; Kalogrides et al., 2013). There are other circumstances in which teachers might be given a rating because of factors that are beyond their control. For example, in most states and districts observation ratings are

not adjusted for grade-level differences because of the assumption that observation results are comparable across grade levels. Mihaly and McCaffrey (2014) used data from grade 4–9 math and English language arts teachers to test the relationship between ratings on three observation protocols (FFT, CLASS, and PLATO) and grade levels. Their regression analysis revealed that middle school teachers score lower than elementary school teachers in all domains on all three observation protocols, that the ratings for elementary school teachers exhibited greater variation, and that the grade-level differences in observation ratings could not be accounted for by differences in teacher, classroom, school, or rater characteristics. They offered several possible explanations for the findings:

- Middle school teachers are truly less effective.
- Systematic differences in the versions of the protocols exist across grade levels.
- Students do not exhibit equal classroom behavior across grade levels because of different developmental stages.
- Teachers respond differently to the protocols across grade levels.

If the last three of these potential explanations are true, using observation ratings that have not been adjusted for grade-level differences would yield less trustworthy results.

Lazarev and Newman (2013) also documented a relationship between observation ratings and grade level. They found that the shape of the graph of the relationships between observation ratings and teacher value-added tends to be nonlinear and to differ between elementary and middle school.

Chaplin et al. (2014) also investigated correlations between observation ratings and other student and school characteristics. Teacher ratings were negatively correlated with the percentage of students eligible for the federal school lunch program and the percentage of racial/ethnic minority students but were positively correlated with the percentage of students designated as gifted. No correlation was found with the percentage of English learner students, students in special education, or female students. A closer examination of these facets of variability would provide a stronger basis for making observations a useful part of the teacher evaluation system.

## Appendix B. Texas Teacher Evaluation and Support System

The primary focus of the Texas Teacher Evaluation and Support System (T-TESS) is to provide continuous formative feedback for teachers to improve their teaching practices and ultimately to improve student outcomes.[9] The steering committee for the development of T-TESS, which comprised teachers, principals, and representatives from higher education and educator organizations, worked from fall 2013 through spring 2014 to update the Texas Teacher Standards (Texas Administrative Code, Chapter 149, Section 149.1001) and to build a standards-aligned rubric. The evaluation system requires teachers to develop goal-setting and professional development plans and to undergo an evaluation cycle that includes preconferences, observations, and postconferences. During the pilot two instruments were used to measure teacher effectiveness: a rubric with which teachers were rated on 16 dimensions across four domains (planning, instruction, learning environment, and professional practices and responsibilities), and a student growth measure.

Districts are able to report each of the 16 dimension ratings from the rubric and the student growth measure separately. For districts that report an aggregate score, the Texas Education Agency recommends a weighting of 80 percent for ratings from the rubric and 20 percent for measures of student growth. The agency found that the option to report disaggregated ratings has led to greater integrity in the rating process because raters are not concerned with how dimension ratings are aggregated and thus can be more forthright in assigning dimension ratings (Tim Regal, director of educator evaluation and support at the Texas Education Agency, personal communication, June 10, 2016).

### Rubric

The T-TESS rubric comprises 16 dimensions across four domains: planning, instruction, learning environment, and professional practices and responsibilities (table B1). Teachers are assigned a rating (improvement needed, developing, proficient, accomplished, or distinguished) for each dimension. According to the Texas Education Agency, during the pilot year, domain ratings were usually obtained by averaging dimension ratings and converting that average into a performance level using the aggregation method presented in box 2 in the main text. Domain ratings were also usually averaged to arrive at an overall rating.[10] Districts were not required or encouraged to convert ordinal ratings into numeric values at the overall, domain, or dimension levels.

Near the beginning of the school year teachers conducted a self-assessment by reviewing teacher and student data to formulate their professional growth goals and plan. Teachers then met with a rater at a goal-setting conference to review and adjust the goals as needed. Throughout the year teachers monitored their progress toward the professional development goals. Several weeks before the school year ended, teachers and their raters gathered data collected throughout the year to discuss at the end-of-year conference and to formulate tentative goals for the following school year. The ratings for this goal-setting and professional development plan were embedded in dimensions 4.2 (goal setting) and 4.3 (professional development). As such, the first three domains were rated from evidence collected during preconferences and classroom observations. The fourth domain was scored after the teacher and rater discussed evidence related to each of the four dimensions (for example, demeanor and ethics, goal setting, professional development, and school community involvement) at the end-of-year conference (Texas Education Agency, 2016a).

**Table B1. Domains and dimensions on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot**

| Domain and dimension | Description |
|---|---|
| Domain 1: Planning | |
| 1.1 Standards and alignment | The teacher designs clear, well-organized sequential lessons that reflect best practice, align with standards, and are appropriate for diverse learners. |
| 1.2 Data and assessment | The teacher uses formal and informal methods to measure student progress, then manages and analyzes student data to inform instruction. |
| 1.3 Knowledge of students | Through knowledge of students and proven practices, the teacher ensures high levels of learning, social-emotional development, and achievement for all students. |
| 1.4 Activities | The teacher plans engaging, flexible lessons that encourage higher-order thinking, persistence, and achievement. |
| Domain 2: Instruction | |
| 2.1 Achieving expectations | The teacher supports all learners in their pursuit of high levels of academic and social-emotional success. |
| 2.2 Content knowledge and expertise | The teacher uses content and pedagogical expertise to design and execute lessons aligned with state standards, related content, and student needs. |
| 2.3 Communication | The teacher clearly and accurately communicates to support persistence, deeper learning, and effective effort. |
| 2.4 Differentiation | The teacher differentiates instruction, aligning methods and techniques to diverse student needs. |
| 2.5 Monitor and adjust | The teacher formally and informally collects, analyzes, and uses student progress data and makes needed lesson adjustments. |
| Domain 3: Learning environment | |
| 3.1 Classroom environment, routines, and procedures | The teacher organizes a safe, accessible, and efficient classroom. |
| 3.2 Managing student behavior | The teacher establishes, communicates, and maintains clear expectations for student behavior. |
| 3.3 Classroom culture | The teacher leads a mutually respectful and collaborative class of actively engaged learners. |
| Domain 4: Professional practices and responsibilities | |
| 4.1 Professional demeanor and ethics | The teacher meets district expectations for attendance, professional appearance, and decorum, procedural, ethical, legal and statutory responsibilities. |
| 4.2 Goal setting | The teacher reflects on his/her practice. |
| 4.3 Professional development | The teacher enhances the professional community. |
| 4.4 School community involvement | The teacher demonstrates leadership with students, colleagues, and community members in the school, district, and community through effective communication and outreach. |

**Note:** During the refinement phase in 2015/16, several minor changes were made to the descriptions that were used in the pilot (Tim Regal, director of educator evaluation and support at the Texas Education Agency, personal communication, June 10, 2016).

**Source:** Texas Education Agency, 2016a.

Teachers were observed by qualified raters, including administrators, teacher leaders, and district personnel. To become qualified during the 2014/15 pilot year, raters had to attend a two-day face-to-face training and a one-day online training and demonstrate proficiency in observation appraisal by completing an online assessment that included scoring a lesson and postconference responses. The face-to-face component of the training focused on the observation cycle and placed little emphasis on the goal-setting and professional development plan at the end-of-year conference protocol.[11] Raters were required to complete subsequent certifications to remain current. During the pilot raters appraised teachers in informal walkthroughs or in announced or unannounced formal sessions; they then entered the rubric ratings into the National Institute of Education and Training's online system (Texas Education Agency, 2016a). Ratings from formal walkthroughs and observations, either announced or unannounced, were counted in the final ratings (Tim Regal, director of educator evaluation and support at the Texas Education Agency, personal communication, March 22, 2016).

### Student growth measures

The student growth component measures the academic progress that students make during their time with a particular teacher. Districts have the authority to choose any measures for any given grade or subject, and value-added measures do not necessarily have to be used for teachers in tested subjects. During the pilot some student growth measures were piloted for informational purposes in select districts. Value-added data were provided to T-TESS pilot districts during the pilot (2014/15) but were not widely provided to districts during the refinement phase (2015/16). No student growth data were used in calculating teachers' overall ratings during these two years. Districts that plan to use value-added data in 2017/18 and beyond are advised to produce and fund the measure on their own (Texas Education Agency, 2016b). Student growth data from the 2014/15 pilot were not available to the study team at the time of analysis, and the study team did not have access to student growth data from any other year.

## Appendix C. Data and methods

This appendix presents a detailed description of the data and methods used in this study. The data section includes information about data sources and the process for transforming the raw data into the analytic file. The methods section specifies the methods used to analyze the data for each of the research questions.

### Data

Two sets of data were used to answer the study's four research questions: teacher-level ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System (T-TESS) pilot (rubric ratings), which were provided by the Texas Education Agency (TEA), and school characteristics, which were from the Texas Academic Performance Review (TAPR) database.

***Texas Teacher Evaluation and Support System pilot rubric ratings.*** During the pilot year raters entered rubric ratings into an online system managed by the National Institute of Excellence in Teaching. That dataset was provided to the Texas Education Agency as deidentified data, which were then shared with the study team. The study team received two data files containing the T-TESS pilot rubric ratings from the Texas Education Agency, one in December 2015 and the other in April 2016. The two files contained the rubric ratings for the same sample of teachers who participated in the pilot but differed by a few data elements. The first data file had unique teacher identification numbers (IDs) generated by the National Institute of Excellence in Teaching, which made it possible to identify teachers who had multiple observations but which did not include identifiable school IDs. The second file had school and district names but no unique teacher IDs. To produce a data file with both the unique teacher IDs and the school IDs (so that rubric ratings could later be linked to school characteristics), the study team merged the two data files of rubric ratings using the following data elements: grade level; date and start and end time of observation; whether the observation was announced or unannounced; rater type; and the dimension, domain, and overall T-TESS rubric ratings. The merged dataset had 11,541 records.

The Texas Education Agency advised the study team that only rubric ratings from formal walkthroughs and observations, announced or unannounced, should be retained. This reduced the number of records from 11,541 to 9,190. The majority of teachers (7,315) had just one record. For technical reasons, domain 4 ratings were entered into separate records for some teachers, in which case the study team combined the records into one. In a few cases of multiple complete observations, only the most recent observation was retained. Five teachers did not have rubric ratings and were removed from the dataset. The final dataset had 8,250 records of rubric ratings across 251 schools and 51 districts.

For each record, rubric ratings at the dimension, domain, and overall levels were provided to the study team as ordinal ratings (improvement needed, developing, proficient, accomplished, or distinguished). According to the Texas Education Agency, during the pilot year districts had the option to convert each dimension rating into a score on a five-point numerical scale (from 1 = improvement needed to 5 = distinguished). Most domain ratings were generated by averaging the dimension scores in each domain to calculate the domain score, but 2 percent of domain ratings appeared to have been made separately by the raters

(for example, a teacher received dimension ratings of 4, 4, 5, and 5, which should have resulted in a domain score of 4.5, accomplished, but he or she received a domain rating other than accomplished). Overall scores were the average of the domain scores. Overall and domain scores were converted to an ordinal rating: scores below 2.0 became a rating of improvement needed, scores of 2.0–2.99 became a rating of developing, scores of 3.0–3.99 became a rating of proficient, scores of 4.0–4.99 became a rating of accomplished, and scores of 5.0 became a rating of distinguished. This conversion was, effectively, a rounding down of the average to the nearest integer. For example, if the average of the dimension ratings was a 4.7, the domain rating assigned would be accomplished, rather than distinguished had 4.7 been rounded up. The study team used the data (ordinal ratings) as they were provided without recalculating any domain ratings on the basis of dimension scores.

***School characteristics data.*** After merging the rubric ratings with data on school characteristics from the Texas Academic Performance Review database, the study team removed 175 records that were not able to be matched with a school.[12] The resulting dataset that was used for research question 4 consisted of 8,080 records of teacher rubric ratings across 251 schools and 51 districts.

Data that were obtained from the Texas Academic Performance Review database included the following characteristics for the 2014/15 school year:

- General profile information: grade span, school locale (rural, town, suburban, urban), and number of students.
- Demographic information: racial/ethnic distribution of students, percentage students eligible for the federal school lunch program, percentage of students who are English learner students, and percentage of students in special education.
- Achievement information: percentage of students who receive at least a satisfactory rating on the State of Texas Assessments of Academic Readiness reading test for grades 3–8 and distinction designation received for academic achievement in English language arts/reading and math.[13]
- Teacher information: percentages of teachers with various years of experience[14] and percentage of teachers with a master's or doctoral degree.

### Methods

The study employed descriptive analysis, correlational analysis, factor analysis, and linear regression to address the research questions.

Research question 1 examined the extent to which the overall, domain, and dimension ratings on the T-TESS rubric differentiate teacher effectiveness. The study team conducted descriptive analysis and reported score means, standard deviations, and the percentage of teachers receiving each rating. At the dimension level the percentage of teachers receiving the modal score was reported.

Research question 2 assessed the internal consistency of the T-TESS rubric. The study team calculated pairwise correlations between the four domain ratings and between 16 dimension ratings. Positive and statistically significant values of the pairwise correlations would have indicated that dimensions and domains were mutually consistent. Negative values or low (not statistically significant) values of the correlation coefficient would have indicated that some elements were not measuring the same concept of teacher effectiveness

and thus could not have been meaningfully aggregated and interpreted. The system would be considered to be internally inconsistent.

Research question 3 assessed the extent to which the T-TESS rubric is efficient. The study team conducted an exploratory factor analysis to establish uniqueness of each dimension. (The results of the exploratory factor analysis were used to select the number of factors: the group of dimensions that may have been related to certain hypothetical latent aspects of teacher effectiveness; see appendix F.) The study team examined uniqueness of dimensions to find whether any dimension adds little or no uniqueness to a certain factor. Uniqueness is a statistical metric that can be produced by factor analysis and represents the proportion of variance of a variable or dimension that cannot be attributed to any other variables or dimensions in the model (that is, unique to the variable). The uniqueness value is a single number for each dimension that ranges from 0, fully correlated with other dimensions already measured in the rubric and therefore redundant, to 1, not at all correlated with other dimensions (Cattell, 1973; Kim & Mueller, 1978).

Research question 4 examined how rubric ratings related to school-level characteristics. First, three tests were conducted (*t*-test, Wilcoxon test, and the nonparametric Kolmogorov-Smirnov test) to identify any statistically significant differences in rubric ratings between teachers of different subgroups. The study team created pairs of subgroups based on the bottom and top quintile of percentage of the following:

- Racial/ethnic minority students.
- Students eligible for the federal school lunch program.
- English learner students.
- Students in special education.
- Teachers with five or fewer years of experience.

The differences in rubric ratings were also compared between the most advantaged schools (those in the bottom quintile of percentage of students eligible for the federal school lunch program and with math and reading academic distinction) and disadvantaged schools (those in the top quintile of students eligible for the federal school lunch program and with no academic distinction).

This approach, with some differences in how variables were created, followed that of Goldhaber, Walch, and Gabele's (2012) in their study of how model choice could affect teacher evaluation results. Goldhaber et al.'s approach involved three groups—advantaged, average, and disadvantaged classrooms—which were based on the aggregate student-level average prior achievement (an average of math and reading test scores) and the percentage of students in the classroom eligible for the federal school lunch program. The current study had a binary classification of schools—advantaged or disadvantaged—and academic distinctions were used in lieu of prior achievement because prior achievement data were not available.

The study team then estimated models of the following structure, one for the overall rubric rating and one for each of the four domain ratings: $M_i = \alpha + X_i\beta + \varepsilon_i$ where $M_i$ is teacher $i$'s domain or dimension score, $\alpha$ is the constant term, $X_i$ is the vector of school characteristics, and $\varepsilon_i$ is the teacher-level error term.

The models were estimated for the full sample, as well as for each of the school grade-level (elementary, middle, and high school) subsamples. For the full sample the results for two models were presented. The first model included variables that have a theoretical basis or have been shown in existing studies of teacher evaluation systems to possibly relate to teacher evaluation ratings. The second model was a reduced model, which was obtained by repeatedly removing the least significant terms and comparing the levels of information criterion (Akaike information criterion) for the sequential models. The resulting model was the most efficient model: it was the best combination of parsimony (number of covariates) and explanatory power (explained variance).

## Appendix D. Comparison between characteristics of Texas Teacher Evaluation and Support System pilot districts and all Texas districts

This appendix compares the demographic and locale characteristics of districts that participated in the 2014/15 Texas Teacher Evaluation and Support System (T-TESS) pilot and all districts in the state.

For most demographic characteristics T-TESS pilot districts show no statistically significant differences from the state average (table D1). The only exceptions are that T-TESS pilot districts average more teachers per school and have a lower percentage of middle schools.

The distribution of pilot districts by locale type, however, differs significantly from that of the rest of the state; a Pearson's chi-squared test resulted in a *p*-value of 0.018 (table D2). The difference may be driven by the larger percentages of town and suburban districts in the pilot sample.

**Table D1. Comparison of demographic characteristics between 2014/15 Texas Teacher Evaluation and Support System pilot districts and all Texas districts**

| | Pilot districts ($n$ = 59) | | | All Texas districts ($n$ = 1,240) | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic | Minimum | Maximum | Mean | Minimum | Maximum | Mean | Difference | *p* value |
| Number of schools | 1 | 78 | 9.32 | 1 | 288 | 7.53 | 1.79 | 0.28 |
| Teachers per school | 6.53 | 53.95 | 28.05 | 0.48 | 101.80 | 24.13 | 3.92 | 0.02* |
| Students per school | 68 | 849.20 | 393.40 | 13 | 1,732 | 338.80 | 54.60 | 0.05 |
| Percentage of students eligible for the federal school lunch program | 7 | 98 | 61.64 | 0 | 99 | 59.39 | 2.25 | 0.39 |
| Percentage of English learner students | 0 | 54 | 10.34 | 0 | 92 | 8.82 | 1.51 | 0.35 |
| Percentage of students in special education | 3 | 21 | 9.20 | 0 | 99 | 9.20 | 0.00 | 0.99 |
| Percentage of elementary schools | 20 | 100 | 44.93 | 13 | 100 | 45.95 | –1.02 | 0.65 |
| Percentage of middle schools | 6 | 40 | 24.57 | 5 | 100 | 27.12 | –2.55 | 0.04* |
| Percentage of high schools | 4 | 50 | 26.67 | 4 | 100 | 28.72 | –2.05 | 0.24 |
| Race/ethnicity | | | | | | | | |
| Percentage of students who are American Indian | 0 | 2 | 0.31 | 0 | 22 | 0.40 | –0.09 | 0.15 |
| Percentage of students who are Asian | 0 | 16 | 1.48 | 0 | 48 | 1.32 | 0.16 | 0.68 |
| Percentage of students who are Black | 0 | 77 | 9.49 | 0 | 99 | 9.89 | –0.4 | 0.84 |
| Percentage of students who are Hispanic | 4 | 98 | 46.29 | 1 | 100 | 40.13 | 6.16 | 0.08 |
| Percentage of students who are White | 1 | 93 | 40.51 | 0 | 96 | 46.24 | –5.73 | 0.12 |

* Significant at $p < .05$.

**Source:** U.S. Department of Education, 2013.

**Table D2. Comparison by locale composition between 2014/15 Texas Teacher Evaluation and Support System pilot districts and all Texas districts**

| | Pilot districts (*n* = 59) | | All Texas districts (*n* = 1,240) | |
|---|---|---|---|---|
| Locale | Number | Share of total (percent) | Number | Share of total (percent) |
| Town | 17 | 29 | 220 | 18 |
| Urban | 7 | 12 | 230 | 19 |
| Rural | 24 | 41 | 647 | 52 |
| Suburb | 11 | 19 | 143 | 12 |

**Note:** A Pearson's chi-squared test results in a *p*-value of 0.018, indicating that the distribution of pilot districts by locale differs significantly from that of all Texas districts. Percentages may not sum to 100 because of rounding.

**Source:** U.S. Department of Education, 2013.

## Appendix E. Detailed results

This appendix provides the detailed results of the analysis of the Texas Teacher Evaluation and Support System (T-TESS) pilot data. Table E1 presents the correlations between dimension ratings on the T-TESS rubric (research question 2). Table E2 presents the uniqueness values for each of the dimension ratings on the T-TESS rubric for a two-factor, three-factor, and four-factor model (research question 3). Table E3 shows a comparison of overall rubric ratings between schools in the top quintile and those in the bottom quintile for the school characteristics included in the analysis. Table E4 shows a comparison of the overall rubric rating between advantaged and disadvantaged schools. Tables E5 and E6 present similar comparisons for domain ratings. Tables E7–E10 present a comparison of regression results for several models; models 1 and 2 use the full sample, and models 3–5 use the elementary, middle, and high school subsamples (research question 4). The dependent variable in each is one of the domain ratings. Table E11 provides the descriptive statistics for characteristics of the schools that participated in the 2014/15 T-TESS pilot.

**Table E1. Correlations between dimension ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot**

| | Dimension | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dimension** | **1.1** | **1.2** | **1.3** | **1.4** | **2.1** | **2.2** | **2.3** | **2.4** | **2.5** | **3.1** | **3.2** | **3.3** | **4.1** | **4.2** | **4.3** |
| Domain 1: Planning | | | | | | | | | | | | | | | |
| 1.2 | .53 | | | | | | | | | | | | | | |
| 1.3 | .50 | .53 | | | | | | | | | | | | | |
| 1.4 | .51 | .47 | .51 | | | | | | | | | | | | |
| Domain 2: Instruction | | | | | | | | | | | | | | | |
| 2.1 | .51 | .50 | .54 | .55 | | | | | | | | | | | |
| 2.2 | .55 | .49 | .54 | .54 | .53 | | | | | | | | | | |
| 2.3 | .49 | .48 | .53 | .54 | .54 | .53 | | | | | | | | | |
| 2.4 | .48 | .52 | .54 | .54 | .51 | .48 | .48 | | | | | | | | |
| 2.5 | .49 | .54 | .55 | .52 | .54 | .50 | .53 | .58 | | | | | | | |
| Domain 3: Learning environment | | | | | | | | | | | | | | | |
| 3.1 | .47 | .46 | .50 | .50 | .53 | .48 | .52 | .49 | .51 | | | | | | |
| 3.2 | .46 | .46 | .51 | .45 | .49 | .46 | .49 | .49 | .53 | .65 | | | | | |
| 3.3 | .47 | .47 | .54 | .51 | .53 | .5 | .53 | .52 | .54 | .62 | .63 | | | | |
| Domain 4: Professional practices and responsibilities | | | | | | | | | | | | | | | |
| 4.1 | .41 | .42 | .43 | .38 | .40 | .43 | .39 | .37 | .38 | .40 | .42 | .43 | | | |
| 4.2 | .43 | .48 | .45 | .41 | .42 | .42 | .43 | .42 | .41 | .40 | .41 | .44 | .52 | | |
| 4.3 | .39 | .43 | .40 | .37 | .36 | .39 | .38 | .36 | .37 | .37 | .37 | .41 | .49 | .54 | |
| 4.4 | .37 | .39 | .41 | .36 | .37 | .39 | .38 | .36 | .37 | .36 | .36 | .40 | .51 | .50 | .55 |

**Note:** All correlation coefficients are statistically significant at $p < .001$. Shading represents correlations between dimensions within the same domain.

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E2. Uniqueness values for two-factor, three-factor, and four-factor models of dimension ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot**

| Domain and dimension | Two factor model | Three factor model | Four factor model |
|---|---|---|---|
| Domain 1: Planning | | | |
| 1.1 Standards and alignment | .53 | .51 | .50 |
| 1.2 Data and assessment | .52 | .50 | .48 |
| 1.3 Knowledge of students | .46 | .46 | .46 |
| 1.4 Activities | .49 | .46 | .46 |
| Domain 2: Instruction | | | |
| 2.1 Achieving expectations | .47 | .46 | .45 |
| 2.2 Content knowledge and expertise | .50 | .48 | .43 |
| 2.3 Communication | .49 | .48 | .48 |
| 2.4 Differentiation | .49 | .49 | .43 |
| 2.5 Monitor and adjust | .45 | .45 | .41 |
| Domain 3: Learning environment | | | |
| 3.1 Classroom environment, routines, and procedures | .46 | .37 | .35 |
| 3.2 Managing student behavior | .48 | .33 | .33 |
| 3.3 Classroom culture | .43 | .38 | .38 |
| Domain 4: Professional practices and responsibilities | | | |
| 4.1 Professional demeanor and ethics | .51 | .51 | .50 |
| 4.2 Goal setting | .47 | .47 | .47 |
| 4.3 Professional development | .45 | .46 | .46 |
| 4.4 School community involvement | .48 | .48 | .48 |

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E3. Average overall rubric ratings from the 2014/15 Texas Teacher Evaluation and Support System pilot for teachers in schools in the top quintile and teachers in schools in the bottom quintile of school characteristics**

| Quintile of school characteristics | Percentage of students eligible for the federal school lunch program | Percentage of racial/ethnic minority students | Percentage of English learner students | Percentage of students in special education | Percentage of teachers with five or fewer years of experience |
|---|---|---|---|---|---|
| Bottom quintile | 2.91 (0.61) | 2.81 (0.61) | 2.85 (0.63) | 2.84 (0.63) | 2.81 (0.53) |
| Top quintile | 2.78 (0.58) | 2.76 (0.59) | 2.80 (0.62) | 2.73 (0.58) | 2.75 (0.65) |
| Difference in average overall rating | 0.13 | 0.05 | 0.05 | 0.11 | 0.06 |

**Note:** Numbers in parentheses are standard deviations. Each subcategory has approximately 48–50 schools (out of 251 total schools). All differences were statistically significant at $p < .001$.

**Source:** Authors' calculations based on data for the 2014/15 school year from the Texas Academic Performance Report database and on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

## Table E4. Average overall rubric ratings from the 2014/15 Texas Teacher Evaluation and Support System pilot for teachers in advantaged and disadvantaged schools

| School group | Average overall rating |
|---|---|
| Advantaged school | 2.89 (0.67) |
| Disadvantaged school | 2.79 (0.67) |
| Difference in average overall rating | 0.10 |

**Note:** Numbers in parentheses are standard deviations. Advantaged schools are schools that are in the bottom quintile of percentage of students eligible for the federal school lunch program and that received an academic distinction designation in English language arts/reading and math ($n$ = 19); disadvantaged schools are schools that are in the top quintile of percentage of students eligible for the school lunch program and that did not receive an academic distinction designation ($n$ = 26). All differences were statistically significant at $p < .001$.

**Source:** Authors' calculations based on data for the 2014/15 school year from the Texas Academic Performance Report database and on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

## Table E5. Average domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot for teachers in schools in the top quintile and teachers in schools in the bottom quintile of school characteristics

| Domain and quintile of school characteristics | Percentage of students eligible for the federal school lunch program | Percentage of racial/ethnic minority students | Percentage of English learner students | Percentage of students in special education | Percentage of teachers with five or fewer years of experience |
|---|---|---|---|---|---|
| Domain 1: Planning | | | | | |
| Bottom quintile | 3.07 (0.61) | 3.00 (0.61) | 3.03 (0.62) | 3.02 (0.59) | 2.96 (0.48) |
| Top quintile | 2.94 (0.54) | 2.93 (0.57) | 2.96 (0.57) | 2.90 (0.54) | 2.92 (0.59) |
| Difference in average domain rating | 0.12*** | 0.07*** | 0.06*** | 0.12*** | 0.05*** |
| Domain 2: Instruction | | | | | |
| Bottom quintile | 3.02 (0.61) | 2.92 (0.60) | 2.98 (0.63) | 2.95 (0.61) | 2.92 (0.52) |
| Top quintile | 2.89 (0.58) | 2.86 (0.57) | 2.90 (0.61) | 2.86 (0.58) | 2.87 (0.64) |
| Difference in average domain rating | 0.14*** | 0.07*** | 0.08*** | 0.09*** | 0.04*** |
| Domain 3: Learning environment | | | | | |
| Bottom quintile | 3.32 (0.71) | 3.21 (0.68) | 3.21 (0.68) | 3.29 (0.73) | 3.16 (0.59) |
| Top quintile | 3.16 (0.68) | 3.19 (0.69) | 3.22 (0.69) | 3.12 (0.67) | 3.16 (0.77) |
| Difference in average domain rating | 0.16*** | 0.02 | –0.01 | 0.17*** | –0.00 |
| Domain 4: Professional practices and responsibilities | | | | | |
| Bottom quintile | 3.20 (0.62) | 3.09 (0.55) | 3.14 (0.63) | 3.15 (0.60) | 3.05 (0.46) |
| Top quintile | 3.04 (0.49) | 3.03 (0.51) | 3.06 (0.54) | 3.00 (0.54) | 3.03 (0.60) |
| Difference in average domain rating | 0.16*** | 0.06*** | 0.08*** | 0.15*** | 0.01 |

*** Significant at $p < .001$.

**Note:** Numbers in parentheses are standard deviations. The number of schools in each subcategory is approximately 48–50 (out of 251 total schools).

**Source:** Authors' calculations based on data for the 2014/15 school year from the Texas Academic Performance Report database and on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E6. Average domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot between teachers in disadvantaged and advantaged schools**

| Domain and school group | Average domain rating |
|---|---|
| Domain 1: Planning | |
| Advantaged school | 3.05<br>(0.66) |
| Disadvantaged school | 2.95<br>(0.63) |
| Difference in average domain rating | 0.10*** |
| Domain 2: Instruction | |
| Advantaged school | 3.02<br>(0.66) |
| Disadvantaged school | 2.89<br>(0.64) |
| Difference in average domain rating | 0.13*** |
| Domain 3: Learning environment | |
| Advantaged school | 3.25<br>(0.73) |
| Disadvantaged school | 3.17<br>(0.75) |
| Difference in average domain rating | 0.07*** |
| Domain 4: Professional practices and responsibilities | |
| Advantaged school | 3.18<br>(0.64) |
| Disadvantaged school | 3.06<br>(0.55) |
| Difference in average domain rating | 0.12*** |

*** Significant at $p < .001$.

**Note:** Numbers in parentheses are standard deviations. Advantaged schools are schools that are in the bottom quintile of percentage of students eligible for the federal school lunch program and that received an academic distinction designation in English language arts/reading and math ($n = 19$); disadvantaged schools are schools that are in the top quintile of percentage of students eligible for the school lunch program and that did not receive an academic distinction designation ($n = 26$).

**Source:** Authors' calculations based on data for the 2014/15 school year from the Texas Academic Performance Report database and on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E7. Regression results for the relationship between planning domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot and school characteristics**

| Intercept and covariate | Model 1: full sample, full model | Model 2: full sample, reduced model | Model 3: elementary school sample, full model | Model 4: middle school sample, full model | Model 5: high school sample, full model |
|---|---|---|---|---|---|
| Intercept | 3.119*** (0.289) | 2.872*** (0.205) | 3.086*** (0.444) | 1.872* (0.843) | 3.233*** (0.572) |
| School grade span (high school is reference group) | | | | | |
| Elementary | 0.105 (0.065) | na | na | na | na |
| Middle | 0.042 (0.064) | na | na | na | na |
| School locale (urban is reference group) | | | | | |
| Suburb | –0.085 (0.072) | na | –0.062 (0.097) | –0.226 (0.144) | 0.048 (0.144) |
| Town | –0.092 (0.081) | na | –0.046 (0.109) | –0.198 (0.175) | 0.141 (0.182) |
| Rural | –0.040 (0.08) | na | –0.045 (0.108) | –0.154 (0.16) | 0.207 (0.183) |
| School size (unit of change: 1,000 students) | | | | | |
| Number of students | 0.081 (0.043) | 0.071 (0.037) | 0.091 (0.139) | 0.144 (0.173) | 0.145* (0.062) |
| School demographics (unit of change: 1 percentage point) | | | | | |
| Percentage of students who are White | 0.000 (0.001) | na | 0.002 (0.002) | –0.003 (0.003) | 0.001 (0.003) |
| Percentage of students who are eligible for the federal school lunch program | –0.005* (0.002) | –0.004*** (0.001) | –0.004. (0.002) | 0.000 (0.004) | –0.007 (0.004) |
| Percentage of English learner students | 0.003* (0.001) | 0.004** (0.001) | 0.004* (0.002) | –0.000 (0.004) | 0.008 (0.006) |
| Percentage of students in special education | –0.004 (0.007) | na | –0.001 (0.01) | –0.014 (0.015) | 0.005 (0.015) |
| School achievement (unit of change: 1 percentage point) | | | | | |
| Percentage of students scoring proficient on the State of Texas Assessments of Academic Readiness reading test | 0.001 (0.002) | 0.003 (0.002) | 0.000 (0.003) | 0.016* (0.007) | –0.004 (0.005) |
| School teacher information (unit of change: 1 percentage point) | | | | | |
| Percentage of teachers with five or fewer years of experience | –0.002 (0.001) | na | –0.002 (0.002) | 0.004 (0.003) | –0.001 (0.004) |
| Percentage of teachers with a master's or doctoral degree | 0.002 (0.002) | na | 0.006* (0.002) | –0.004 (0.004) | –0.002 (0.004) |
| Number of teachers | 7,365 | 7,365 | 3,402 | 1,745 | 2,218 |
| Number of schools | 221 | 221 | 120 | 52 | 49 |
| Number of districts | 50 | 50 | 43 | 34 | 34 |
| Akaike information criterion | 11,707 | 11,654 | 4,990 | 3,021 | 3,758 |
| Adjusted R-squared | 0.03 | 0.03 | 0.06 | 0.05 | 0.06 |

* Significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .001$.

na is not applicable.

**Note:** Numbers in parentheses are standard errors.

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E8. Regression results for the relationship between instruction domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot and school characteristics**

| Intercept and covariate | Model 1: full sample, full model | Model 2: full sample, reduced model | Model 3: elementary school sample, full model | Model 4: middle school sample, full model | Model 5: high school sample, full model |
|---|---|---|---|---|---|
| Intercept | 3.123*** (0.304) | 3.184*** (0.077) | 2.763*** (0.46) | 1.307 (0.79) | 3.391*** (0.539) |
| School grade span (high school is reference group) | | | | | |
| Elementary | 0.039 (0.067) | na | na | na | na |
| Middle | 0.010 (0.066) | na | na | na | na |
| School locale (urban is reference group) | | | | | |
| Suburb | –0.042 (0.078) | na | 0.014 (0.097) | –0.303* (0.134) | 0.105 (0.135) |
| Town | –0.053 (0.087) | na | –0.008 (0.111) | –0.298. (0.163) | 0.250 (0.172) |
| Rural | –0.020 (0.085) | na | –0.005 (0.111) | –0.277 (0.149) | 0.307 (0.172) |
| School size (unit of change: 1,000 students) | | | | | |
| Number of students | 0.079 (0.045) | 0.070 (0.037) | 0.095 (0.148) | 0.047 (0.163) | 0.164** (0.058) |
| School demographics (unit of change: 1 percentage point) | | | | | |
| Percentage of students who are White | 0.001 (0.001) | na | 0.003 (0.002) | –0.002 (0.003) | 0.001 (0.002) |
| Percentage of students eligible for the federal school lunch program | –0.005** (0.002) | –0.006*** (0.001) | –0.004 (0.003) | –0.000 (0.004) | –0.008 (0.004) |
| Percentage of English learner students | 0.005** (0.001) | 0.005*** (0.001) | 0.006** (0.002) | 0.005 (0.004) | 0.008 (0.005) |
| Percentage of students in special education | –0.001 (0.007) | na | 0.001 (0.01) | 0.013 (0.014) | –0.001 (0.014) |
| School achievement (unit of change: 1 percentage point) | | | | | |
| Percentage of students scoring proficient on the State of Texas Assessments of Academic Readiness reading test | –0.000 (0.003) | na | 0.001 (0.003) | 0.022** (0.007) | –0.005 (0.005) |
| School teacher information (unit of change: 1 percentage point) | | | | | |
| Percentage of teachers with five or fewer years of experience | –0.002 (0.001) | na | –0.001 (0.002) | 0.001 (0.003) | –0.001 (0.003) |
| Percentage of teachers with a master's or doctoral degree | 0.002 (0.002) | na | 0.007** (0.003) | –0.003 (0.004) | –0.004 (0.004) |
| Number of teachers | 7,387 | 7,387 | 3,404 | 1,764 | 2,219 |
| Number of schools | 222 | 222 | 120 | 53 | 49 |
| Number of districts | 50 | 50 | 43 | 35 | 34 |
| Akaike information criterion | 12,230 | 12,164 | 5,533 | 3,055 | 3,747 |
| Adjusted R-squared | 0.03 | 0.03 | 0.05 | 0.05 | 0.07 |

* Significant at $p < .05$; * significant at $p < .01$; *** significant at $p < .001$.

na is not applicable.

**Note:** Numbers in parentheses are standard errors.

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E9. Regression results for the relationship between learning environment domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot and school characteristics**

| Intercept and covariate | Model 1: full sample, full model | Model 2: full sample, reduced model | Model 3: elementary school sample, full model | Model 4: middle school sample, full model | Model 5: high school sample, full model |
|---|---|---|---|---|---|
| Intercept | 3.486*** (0.346) | 3.471*** (0.085) | 3.523*** (0.567) | 1.991* (0.903) | 2.918*** (0.654) |
| School grade span (high school is reference group) | | | | | |
| Elementary | 0.090 (0.079) | na | na | na | na |
| Middle | 0.019 (0.077) | na | na | na | na |
| School locale (urban is reference group) | | | | | |
| Suburb | 0.010 (0.085) | na | 0.074 (0.12) | –0.342 (0.163) | 0.187 (0.167) |
| Town | –0.005 (0.097) | na | 0.023 (0.137) | –0.246 (0.184) | 0.323 (0.208) |
| Rural | 0.029 (0.095) | na | 0.051 (0.136) | –0.256 (0.169) | 0.379 (0.208) |
| School size (unit of change: 1,000 students) | | | | | |
| Number of students | 0.120* (0.052) | 0.098* (0.042) | 0.113 (0.183) | 0.218 (0.174) | 0.241** (0.07) |
| School demographics (unit of change: 1 percentage point) | | | | | |
| Percentage of students who are White | 0.001 (0.001) | na | 0.002 (0.002) | –0.003 (0.003) | 0.004 (0.003) |
| Percentage of students eligible for the federal school lunch program | –0.007** (0.002) | –0.007*** (0.001) | –0.007* (0.003) | –0.006 (0.004) | –0.002 (0.005) |
| Percentage of English learner students | 0.006*** (0.002) | 0.007*** (0.001) | 0.006** (0.002) | 0.007 (0.004) | 0.010 (0.007) |
| Percentage of students in special education | –0.003 (0.008) | na | –0.007 (0.013) | 0.003 (0.015) | 0.017 (0.017) |
| School achievement (unit of change: 1 percentage point) | | | | | |
| Percentage of students scoring proficient on the State of Texas Assessments of Academic Readiness reading test | –0.002 (0.003) | na | –0.002 (0.004) | 0.019* (0.008) | –0.006 (0.006) |
| School teacher information (unit of change: 1 percentage point) | | | | | |
| Percentage of teachers with five or fewer years of experience | –0.001 (0.002) | na | –0.000 (0.002) | 0.003 (0.003) | 0.002 (0.004) |
| Percentage of teachers with a master's or doctoral degree | 0.001 (0.002) | na | 0.004 (0.003) | –0.004 (0.004) | –0.005 (0.005) |
| Number of teachers | 7,386 | 7,386 | 3,404 | 1,764 | 2,218 |
| Number of schools | 222 | 222 | 120 | 53 | 49 |
| Number of districts | 50 | 50 | 43 | 35 | 34 |
| Akaike information criterion | 14,747 | 14,680 | 6,752 | 3,564 | 4,536 |
| Adjusted *R*-squared | 0.06 | 0.04 | 0.04 | 0.06 | 0.0 |

* Significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .001$.

na is not applicable.

**Note:** Numbers in parentheses are standard errors.

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E10. Regression results for the relationship between professional practices and responsibilities domain ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot and school characteristics**

| Intercept and covariate | Model 1: full sample, full model | Model 2: full sample, reduced model | Model 3: elementary school sample, full model | Model 4: middle school sample, full model | Model 5: high school sample, full model |
|---|---|---|---|---|---|
| Intercept | 3.378*** (0.332) | 3.343*** (0.109) | 3.379*** (0.582) | 2.643*** (0.691) | 2.701*** (0.564) |
| School grade span (high school is reference group) | | | | | |
| Elementary | 0.068 (0.077) | na | na | na | na |
| Middle | –0.013 (0.076) | na | na | na | na |
| School locale (urban is reference group) | | | | | |
| Suburb | –0.054 (0.08) | na | 0.004 (0.122) | –0.280* (0.116) | 0.047 (0.135) |
| Town | –0.023 (0.091) | na | –0.009 (0.138) | –0.150 (0.143) | 0.227 (0.172) |
| Rural | –0.003 (0.09) | na | 0.031 (0.137) | –0.235. (0.129) | 0.249 (0.173) |
| School size (unit of change: 1,000 students) | | | | | |
| Number of students | 0.080 (0.05) | 0.074 (0.042) | 0.068 (0.183) | 0.156 (0.141) | 0.195** (0.06) |
| School demographics (unit of change: 1 percentage point) | | | | | |
| Percentage of students who are White | –0.001 (0.001) | na | –0.000 (0.002) | –0.005* (0.002) | 0.001 (0.003) |
| Percentage of students eligible for the federal school lunch program | –0.005* (0.002) | –0.003** (0.001) | –0.006. (0.003) | –0.003 (0.004) | –0.002 (0.004) |
| Percentage of English learner students | 0.002 (0.002) | na | 0.003 (0.002) | –0.002 (0.003) | 0.002 (0.005) |
| Percentage of students in special education | –0.014. (0.008) | –0.021** (0.007) | –0.008 (0.013) | –0.028* (0.013) | –0.003 (0.014) |
| School achievement (unit of change: 1 percentage point) | | | | | |
| Percentage of students scoring proficient on the State of Texas Assessments of Academic Readiness reading test | 0.001 (0.003) | na | –0.002 (0.004) | 0.014* (0.006) | 0.001 (0.005) |
| Percentage of teachers with five or fewer years of experience | –0.001 (0.002) | na | –0.001 (0.002) | 0.003 (0.002) | 0.003 (0.003) |
| Percentage of teachers with a master's or doctoral degree | 0.004 (0.002) | 0.003 (0.002) | 0.010** (0.003) | –0.007* (0.003) | –0.003 (0.004) |
| Number of teachers | 6,998 | 6,998 | 3,301 | 1,660 | 2,037 |
| Number of schools | 216 | 216 | 117 | 51 | 48 |
| Number of districts | 49 | 49 | 42 | 33 | 33 |
| Akaike information criterion | 10,612 | 10,555 | 4,798 | 2,378 | 3,475 |
| Adjusted *R*–squared | 0.04 | 0.04 | 0.08 | 0.09 | 0.03 |

* Significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .001$.

na is not applicable.

**Note:** Numbers in parentheses are standard errors. Two districts did not provide data for this domain (see appendix C).

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table E11. Descriptive statistics for characteristics of schools that participated in the 2014/15 pilot of the Texas Teacher Evaluation and Support System**

| Characteristic | Number of schools | Mean | Standard deviation |
|---|---|---|---|
| Number of students | 250 | 622.58 | 477.27 |
| Percentage of students who are White | 250 | 30.35 | 28.45 |
| Percentage of students eligible for the federal school lunch program | 250 | 65.13 | 20.54 |
| Percentage of English learner students | 250 | 15.83 | 17.74 |
| Percentage of students in special education | 250 | 9.28 | 6.16 |
| Percentage of students scoring proficient on the State of Texas Assessments of Academic Readiness reading test | 231 | 76.45 | 11.89 |
| Percentage of teachers with 0 years of experience | 250 | 9.82 | 11.47 |
| Percentage of teachers with 1–5 years of experience | 250 | 25.63 | 12.00 |
| Percentage of teachers with a master's degree | 250 | 18.39 | 12.17 |
| Percentage of teachers with a doctoral degree | 250 | 0.47 | 1.44 |

**Source:** Authors' calculations based on data for the 2014/15 school year from the Texas Academic Performance Review database.

## Appendix F. Supplemental analysis: Determining the number of factors from the data of the 2014/15 Texas Teacher Evaluation and Support System pilot

The primary purpose of the exploratory factor analysis was to establish the uniqueness of each dimension in order to examine the extent to which the rubric from the 2014/15 Texas Teacher Evaluation and Support System (T-TESS) pilot is efficient. The results do not change in substantively important ways across two-, three-, or four-factor models (see table E2 in appendix E).

The secondary purpose of the exploratory factor analysis was to identify clusters of dimensions that may be measuring the same latent construct, which is an underlying factor that is not directly observed. This appendix presents the methods used to select the number of factors. For factor analysis the study team used the maximum likelihood estimation method and varimax rotation. The study team also triangulated information from several other approaches and selected a three-factor model. The first factor was defined by dimensions of the planning and instruction domains, the second factor by professional practices and responsibilities, and the third factor by learning environment (table F1). The results imply that the same underlying factor of teacher effectiveness explained both the planning domain and instruction domain.

The first approach was the scree test, which identified the most probable number of factors on the basis of where the steep curve becomes a horizontal line (Cattell, 1966). In this analysis the curve leveled off between two and three factors (figure F1).

The second approach was based on explained variance, which produced results consistent with the scree test. Factors past the third added much less to the explained variance than

**Table F1. Factor loadings for the three-factor model of dimension ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot**

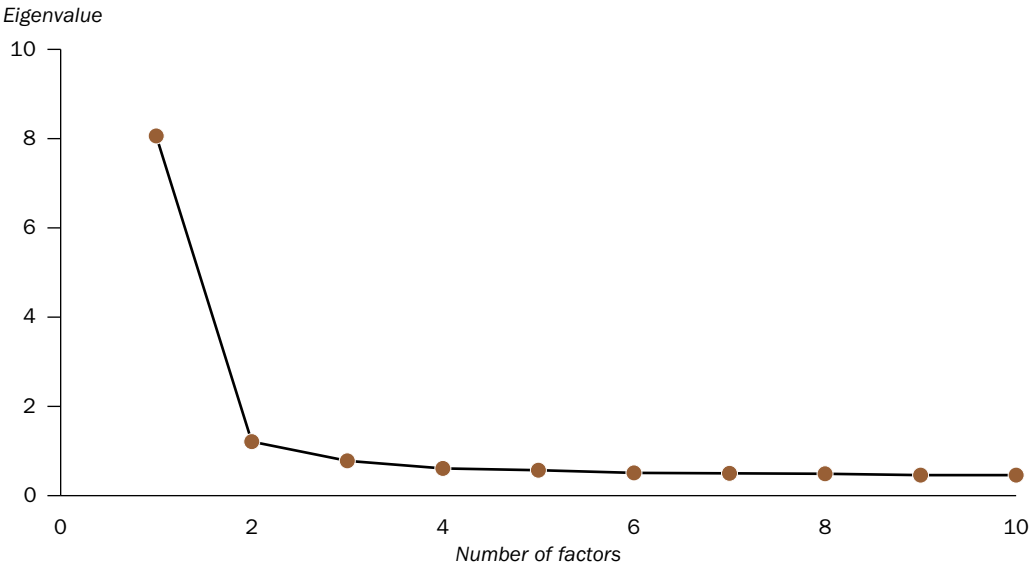
| Dimension | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| 1.1 Standards and alignment | 0.59 | 0.32 | 0.23 |
| 1.2 Data and assessment | 0.56 | 0.37 | 0.23 |
| 1.3 Knowledge of students | 0.59 | 0.32 | 0.31 |
| 1.4 Activities | 0.64 | 0.25 | 0.25 |
| 2.1 Achieving expectations | 0.62 | 0.25 | 0.31 |
| 2.2 Content knowledge and expertise | 0.61 | 0.31 | 0.24 |
| 2.3 Communication | 0.58 | 0.28 | 0.32 |
| 2.4 Differentiation | 0.60 | 0.25 | 0.30 |
| 2.5 Monitor and adjust | 0.61 | 0.25 | 0.35 |
| 3.1 Classroom environment, routines, and procedures | 0.42 | 0.24 | 0.64 |
| 3.2 Managing student behavior | 0.36 | 0.25 | 0.69 |
| 3.3 Classroom culture | 0.45 | 0.30 | 0.57 |
| 4.1 Professional demeanor and ethics | 0.27 | 0.60 | 0.24 |
| 4.2 Goal setting | 0.33 | 0.62 | 0.19 |
| 4.3 Professional development | 0.24 | 0.68 | 0.16 |
| 4.4 School community involvement | 0.23 | 0.67 | 0.16 |

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

the first three factors did. Formally, the first three factors explained approximately $10^{-1}$ of total variance while the fourth and subsequent factors explained approximately $10^{-2}$ or less (table F2).

An additional desirable property of the three-factor model is its consistency with past findings. The results aligned well with the model developed in a study of the Measures of Effective Teaching project data in which the authors performed a factor analysis on 57 variables collected across different instruments and found a three-factor model to be most appropriate (Lazarev & Newman, 2014). The first factor, constructive, was associated with pedagogical techniques and can be mapped to T-TESS's factor 1, which included the planning and instruction domains. The second factor, effective, was associated with student

**Figure F1. Scree plot for dimension ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot**




**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

**Table F2. Explained variance for two-, three-, and four-factor models of the 2014/15 pilot Texas Teacher Evaluation and Support System rubric**

| Variance | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|---|---|---|---|---|
| Two-factor model | | | | |
| Proportion variance | 0.34 | 0.18 | | |
| Cumulative variance | 0.34 | 0.52 | | |
| Three-factor model | | | | |
| Proportion variance | 0.25 | 0.16 | 0.13 | |
| Cumulative variance | 0.25 | 0.42 | 0.55 | |
| Four-factor model | | | | |
| Proportion variance | 0.23 | 0.16 | 0.14 | 0.02 |
| Cumulative variance | 0.23 | 0.39 | 0.53 | 0.56 |

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

achievement and teachers' skills in following procedures and managing student behavior, which can be mapped to T-TESS's factor 3 (learning environment). The third factor was related to student surveys, which were not collected in the T-TESS pilot. The Measures of Effective Teaching project data did not include measures similar to those in T-TESS's factor 2 (professional practices and responsibilities), but the authors suspected that these dimensions, which were measured outside the classroom, such as goal setting and school community involvement, could form a fourth factor in their model.

If a two-factor model had been selected, the model would have been factor 1 (planning, instruction, and learning environment) and factor 2 (professional practices and responsibilities; table F3). To decide whether factor 1 should be disaggregated further, the study team turned to the evidence from Lockwood et al.'s (2015) study, which analyzed data from 450 middle school teachers who participated in the Understanding Teacher Quality study. That study found that there were two underlying factors: quality of instructional practices and quality of teacher classroom management. Lockwood et al.'s finding supported further disaggregating factor 1 in the T-TESS two-factor model into two separate factors, with one factor comprising the planning and instruction domains and the other factor comprising the learning environment domain. The analytic goal of the factor analysis was to find the number of factors that are most replicable. Because the current study examined data from one pilot year of T-TESS and no other studies of T-TESS were available, the study team compared T-TESS's three-factor model to others in the existing body of research. Based on such comparison, coupled with the results of the scree test and analysis of the explained variation in the two-factor, three-factor, and four-factor models, the study team selected a three-factor model.

**Table F3. Factor loadings for the two-factor model of dimension ratings on the rubric from the 2014/15 Texas Teacher Evaluation and Support System pilot**

| Dimension | Factor 1 | Factor 2 |
|---|---|---|
| 1.1 Standards and alignment | 0.59 | 0.34 |
| 1.2 Data and assessment | 0.57 | 0.39 |
| 1.3 Knowledge of students | 0.65 | 0.34 |
| 1.4 Activities | 0.66 | 0.28 |
| 2.1 Achieving expectations | 0.67 | 0.28 |
| 2.2 Content knowledge and expertise | 0.62 | 0.34 |
| 2.3 Communication | 0.65 | 0.30 |
| 2.4 Differentiation | 0.65 | 0.28 |
| 2.5 Monitor and adjust | 0.69 | 0.27 |
| 3.1 Classroom environment, routines, and procedures | 0.69 | 0.26 |
| 3.2 Managing student behavior | 0.67 | 0.27 |
| 3.3 Classroom culture | 0.68 | 0.31 |
| 4.1 Professional demeanor and ethics | 0.34 | 0.61 |
| 4.2 Goal setting | 0.36 | 0.63 |
| 4.3 Professional development | 0.26 | 0.69 |
| 4.4 School community involvement | 0.26 | 0.67 |

**Source:** Authors' calculations based on data from the 2014/15 Texas Teacher Evaluation and Support System pilot provided by the Texas Education Agency.

Because the results of this analysis indicated that an inherent multidimensionality exists in the data of rubric ratings from the T-TESS pilot, a multidimensional approach for presenting the evaluation results that goes beyond aggregating or averaging across all four domains is advantageous.[15] Such an approach can be developed on the basis of discussions among stakeholders about how evaluation ratings should be presented and used to inform decisions for a range of different purposes. For example, while dimension-level ratings may guide strategies for creating a specific teacher's professional development plan, overall ratings may suffice for providing a general sense of how teachers within a school or district are performing. Another direction for expanding understanding of the T-TESS pilot's latent data structure is to investigate how the factor ratings relate to other indicators, such as grade level or incoming achievement ratings, as was explored by Lazarev and Newman (2015). Finally, once a sufficient number of exploratory studies using data collected from T-TESS or a similar rubric are available, a hypothesis of the number of factors could be developed, and confirmatory factor analysis could be conducted to determine whether the results are consistent with the hypothesized number of factors.

## Notes

1. In 2011 the U.S. Department of Education offered states the opportunity to request flexibility waivers for specific requirements of the No Child Left Behind Act. One condition for the waiver was the development of a rigorous and comprehensive system to evaluate and support teacher and principal effectiveness. As of May 2016, 43 states, Puerto Rico, and the District of Columbia had received Elementary and Secondary Education Act flexibility to support higher achievement in schools (U.S. Department of Education, 2016).
2. The REL Southwest Education Effectiveness Research Alliance is a diverse body of approximately 44 stakeholders, including teachers, administrators, researchers, and district and state policymakers. Institutions represented in the alliance include local and state teachers associations, postsecondary institutions, the Texas Education Agency, and other state and district agencies (Regional Educational Laboratory Southwest, n.d.).
3. Although the pilot was conducted in 57 districts, the final analytic sample had 51 districts. See appendix C for a description of the data cleaning process.
4. The difference is statistically significant based on three tests (*t* test, Wilcoxon test, and the nonparametric Kolmogorov-Smirnov test).
5. See table E9 in appendix E for standard deviations of all school characteristics.
6. The study team considered a finding to be substantively important if the increase or decrease in score would have resulted in a shift of one level up or down on the ordinal scale.
7. The pilot data were collected and managed by one entity. For the refinement phase and statewide rollout districts can use their own systems, which makes research beyond the pilot more difficult.
8. Item-level results are ratings on the individual items in the observation protocol.
9. Teach for Texas, https://teachfortexas.org, retrieved June 8, 2016.
10. However, about 2 percent of raters evaluated teachers at both the dimension and domain levels separately, rather than calculating domain ratings by averaging the dimension ratings (see appendix C).
11. During the refinement phase the training was changed to a three-day face-to-face training that heavily emphasized the goal-setting and professional development process measured in domain 4 (Tim Regal, director of educator evaluation and support at the Texas Education Agency, personal communication, June 10, 2016).
12. These 175 records were within the 51 districts in the analytic dataset.
13. The results of the State of Texas Assessments of Academic Readiness for math for grades 3–8 were excluded from the state's accountability system for 2014/15.
14. On the basis of the percentage of total teachers that fell within each range of years of experience, the study team calculated a percentage of teachers with five or fewer years of experience.
15. Districts in Texas have the option to report each of the 16 dimension ratings and the student growth score separately (Tim Regal, director of educator evaluation and support at the Texas Education Agency, personal communication, June 10, 2016).

## References


Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study* (Policy & Practitioner Brief). Seattle, WA: Author. Retrieved July 21, 2016, from http://www.edweek.org/media/17teach-met1.pdf.

Boyd, D., Lankford, H., Loeb, S., Wyckoff, J. (2005). Explaining the short careers of high-achieving teachers in schools with low-performing students. *American Economic Review, 95*(2), 166–171.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Cattell, R. B. (1973). *Factor analysis: An introduction and manual for the psychologist and social scientist*. New York, NY: Harper & Brothers.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools* (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. http://eric.ed.gov/?id=ED545232

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review, 24*(4), 377–392. http://eric.ed.gov/?id=EJ697548

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778–820. http://eric.ed.gov/?id=EJ750956

Cohen, J., & Goldhaber, D. (2016). Observations on evaluating teacher performance: Assessing the strengths and weaknesses of classroom observations and value-added measures. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 8–21). New York, NY : Teachers College Press.

Darling-Hammond, L. (2015). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York, NY: Teachers College Press.

Doherty, K., & Jacobs, S. (2013). *State of the states 2013 connect the dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality. http://eric.ed.gov/?id=ED565882

Elementary and Secondary Education Act of 1965. (1965). Pub. L. No. 89–10, 79 Stat. 27.

Ettema, E., Sengupta, K., & Kress, S. (2014). *A legal lever for enhancing productivity* (Productivity for Results Series No. 3). Dallas, TX: George W. Bush Institute. http://eric.ed.gov/?id=ED560203

Every Student Succeeds Act of 2015. (2015). Pub. L. No. 114–95.

Feng, L. (2010). Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility. *Education Finance and Policy, 5*(3), 278–316. http://eric.ed.gov/?id=EJ892970

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher, 44*(5), 293–307. http://eric.ed.gov/?id=EJ1068119

Goldhaber, D., Walch, J., & Gabele, B. (2012). *Does the model matter? Exploring the relationship between different achievement-based teacher assessments* (CEDR Working Paper No. 2012–6). Seattle, WA: University of Washington.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326–354. http://eric.ed.gov/?id=EJ746489

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance*. Alexandria, VA: Center for Public Education of the National School Boards Association. Retrieved September 1, 2015, from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf.

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103–123. http://eric.ed.gov/?id=EJ998296

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

Kim, J., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. (Sage University Paper series on Quantitative Applications in the Social Sciences, No. 07–014). Newbury Park, CA: Sage.

Kraft, M. A., & Gilmour, A. F. (2016). *Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness*. (Working paper). Providence, RI: Brown University. Retrieved July 21, 2016, from http://scholar.harvard.edu/mkraft/publications/revisiting-widget-effect-teacher-evaluation-reforms-and-distribution-teacher.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62. http://eric.ed.gov/?id=EJ653059

Lash, A., Tran, L., & Huang, M. (2016). *Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system* (REL 2016–135). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. http://eric.ed.gov/?id=ED565904

Lazarev, V., & Newman, D. (2013, September). *How non-linearity and grade-level differences complicate the validation of observation protocols*. Presented at the SREE Fall 2013 Conference, Washington, DC. http://eric.ed.gov/?id=ED563108

Lazarev, V., & Newman, D. (2014, March). *Can multifactor models of teaching improve teacher effectiveness measures?* Presented at the Annual Meeting of the Association for Education Finance and Policy, San Antonio, TX. http://eric.ed.gov/?id=ED558566

Lazarev, V., & Newman, D. (2015, February). *How teacher evaluation is affected by class characteristics: Are observations biased?* Presented at the Annual Meeting of the Association for Education Finance and Policy, Washington, DC. http://eric.ed.gov/?id=ED558567

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. http://eric.ed.gov/?id=ED548027

Lipscomb, S., Chiang, H., & Gill, B. (2012). *Value-added estimates for phase 1 of the Pennsylvania Teacher and Principal Evaluation pilot*. Cambridge, MA: Mathematica Policy Research. http://eric.ed.gov/?id=ED531803

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, 9(3), 1484–1509.

Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 9–49). San Francisco, CA: Jossey-Bass.

No Child Left Behind Act of 2001. (2002). Pub L. No. 107–110, 115 Stat. 1425.

Regional Education Laboratory Southwest. (n.d.). *Educator Effectiveness Research Alliance*. Austin, TX: Author. Retrieved August 17, 2016, from http://relsouthwest.sedl.org/research-alliances/educator_effectiveness.html.

Schultz, S. E., & Pecheone, R. L. (2014). Assessing quality teaching in science. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 444–483). San Francisco, CA: Jossey-Bass.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293–317. http://eric.ed.gov/?id=EJ1100448

Teach for Texas. (n.d.). *T-TESS general info*. Retrieved September 15, 2016, from https://teachfortexas.org/Views/FAQ.

Texas Administrative Code, Chapter 149, Section 149.1001.

Texas Administrative Code, Chapter 150, Section 150.1001.

Texas Education Agency. (2015). *2015 Accountability Manual*. Austin, TX: Author. Retrieved August 3, 2017, from https://rptsvr1.tea.texas.gov/perfreport/account/2015/index.html.

Texas Education Agency. (2016a). *Texas Teacher Evaluation Support System appraiser training handbook*. Austin, TX: Author. Retrieved July 21, 2016, from https://teachfortexas.org/Resource_Files/Guides/T-TESS_Appraiser_Handbook.pdf.

Texas Education Agency. (2016b). *Student growth overview*. Austin, TX: Author. Retrieved December 6, 2016 from http://tea.texas.gov/Texas_Educators/Educator_Evaluation_and_Support_System/Texas_Teacher_Evaluation_and_Support_System/.

Texas Education Agency. (2016c). *Texas Teacher Evaluation and Support System FAQ*. Austin, TX: Author. Retrieved December 6, 2016, from http://www.nctq.org/docs/Texas_Teacher_Evaluation_and_Support_System_FAQ_June_5_74815.pdf.

The New Teacher Project. (2010). *Teacher evaluation 2.0*. Brooklyn, NY: Author. http://eric.ed.gov/?id=ED518129

U.S. Department of Education, National Center for Education Statistics. (2013). Common Core of Data. Public Elementary/Secondary School Universe Survey, 2013–14, v.2a. Retrieved February 17, 2017, from https://nces.ed.gov/ccd/pubschuniv.asp.

U.S. Department of Education. (2016). *ESEA flexibility*. Washington, DC: Author. Retrieved August 3, 2016, from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. 2nd ed. Brooklyn, NY: New Teacher Project. Retrieved July 21, 2016, from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings. http://eric.ed.gov/?id=ED553815

Yoon, P. S., Chen, J., & Holtzman, S. L. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 383–414). San Francisco, CA: Jossey-Bass.

## 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