Making Connections

May 2016

Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system

Andrea Lash
Loan Tran
Min Huang
WestEd

Key findings

This validation study examined principals' evaluation ratings of teachers made on an instrument adapted from the Danielson Framework for Teaching and used in the Washoe County School District in Reno, Nevada in 2012/13. Principals used a four-point rating scale to rate teachers on 22 teaching components. The teaching components were expected to measure four different dimensions of teaching. The analysis showed that principals discriminated among teachers they thought to be effective and highly effective, the two highest points on the rating scale; they rarely identified teachers as minimally effective or ineffective (approximately 10 percent of ratings were in the lowest two categories). Additionally, individual component ratings did not appear to measure distinct aspects of teaching. Instead, the analyses support using an average rating taken over all components. This average rating shows a moderate relationship with student learning, providing some evidence that it may be interpreted as an indicator of teacher effectiveness in promoting learning.

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

REL WEST
Regional Educational Laboratory
At WestEd

U.S. Department of Education
John B. King, Jr., *Secretary*

Institute of Education Sciences
Ruth Neild, *Deputy Director for Policy and Research*
Delegated Duties of the Director

National Center for Education Evaluation and Regional Assistance
Joy Lesnick, *Acting Commissioner*
Amy Johnson, *Action Editor*
OK-Choon Park, *Project Officer*

REL 2016–135

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.

May 2016

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

Lash, A., Tran, L., and 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. Retrieved from http://ies.ed.gov/ncee/edlabs.

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

Summary

States and districts across the country are seeking to improve teacher evaluation systems; many assess teachers based on their teaching practices inside and outside the classroom. This study examines the validity of ratings from a classroom observation instrument in a teacher evaluation system adapted from the Danielson Framework for Teaching[1] (Danielson, 2007) and proposed for use in Washoe County School District in Reno, Nevada. While the study takes place in one school district, the findings may be of interest to districts and states that are using or considering using the Danielson Framework. Previous studies have examined only a portion of the Danielson Framework, typically focusing on the ratings of teachers' classroom practices; this study is unique in examining the full Danielson Framework, including teachers' planning for instruction and carrying out of professional responsibilities.

Examining validity requires identifying and assessing the plausibility of the assumptions that underlie the planned use of the results. This study examined four assumptions underlying Washoe County School District's proposed interpretation of ratings from the classroom observation instrument:

- Ratings from the classroom observation instrument differentiate among teachers.
- Each of the four domain ratings measures a single, cohesive area of teaching practice.
- Each of the four domains is distinct from the others.
- Ratings from the classroom observation instrument indicate teacher effectiveness in promoting student learning.

The assumptions were identified through logical analysis of the district's proposed uses of the observation data. Some of the assumptions were supported by the analyses, and others were not.

The analyses of ratings of 713 teachers by their principal in 2012/13 showed that ratings for the 22 components across the four domains differentiated only among teachers rated as effective or highly effective and identified few teachers as minimally effective or ineffective; on each component nearly all teachers were rated as effective or highly effective. The distributions of the average rating in each domain (calculated as the average rating for the components of each domain) and of the average rating over all 22 components were similar: in all cases at least 80 percent of teachers received an average rating at or above effective.

While component ratings within each of the four domains appeared to measure a single, cohesive trait within the domain, the analyses found that the four domains were not distinct from one another. Instead, the ratings of the 22 components from across the four domains seemed to measure the same cohesive trait.

Although there was no evidence to distinguish one domain from another, each domain rating and the average of all the component ratings across domains correlated positively with student learning in reading and in math, as would be expected if the ratings measured teacher effectiveness in promoting student learning.

Contents

Summary i

Why this study? 1

What the study examined 2
What are the distributions of teacher ratings for the 22 components and four domains? 4
What is the internal consistency of ratings within each domain and across all domains? 4
Do the empirical data support the hypothesized grouping of components into domains? 5
Do ratings from the classroom observation instrument predict student learning? 6

What the study found 6
On each component most teachers were rated as effective or highly effective 6
The component ratings are consistent in the information they provide about teachers 10
The empirical data do not support grouping ratings into four domains 11
Teachers' scores on the classroom observation instrument are related to measures of their students' learning 12

Implications of the study 14

Limitations of the study findings 15

Appendix A. Supporting tables A-1

Appendix B. Confirmatory factor analyses: Methods and findings B-1

Notes Notes-1

References Ref-1

Boxes
1 About the Danielson Framework for Teaching and the Washoe County School District–adapted system 1
2 Study data 3

Figures
1 Percentage of teachers by rating for each component in domain 1, Planning and Preparation 7
2 Percentage of teachers by rating for each component in domain 2, Classroom Environment 7
3 Percentage of teachers by rating for each component in domain 3, Instruction 8
4 Percentage of teachers by rating for each component in domain 4, Professional Responsibilities 8
5 Teachers' averages over all component ratings 9

Tables
1 Internal consistency for 2012/13 ratings by domain 11
2 Pearson correlation coefficients of domain ratings for 2012/13 with student growth, by subject 13

A1 Comparison of domain 1, Planning and Preparation, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system A-1
A2 Comparison of domain 2, Classroom Environment, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system A-2
A3 Comparison of domain 3, Instruction, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system A-3
A4 Comparison of domain 4, Professional Responsibilities, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system A-4
A5 Distribution of 713 teacher ratings, by domain and component A-5
A6 Distribution of 713 teachers' average ratings for four domains A-5
B1 Confirmatory factor analysis of one- and two-factor models B-1
B2 Intercorrelations of factor scores from a four-factor solution of observation ratings B-2

Why this study?

States and districts across the country are developing new systems to evaluate teachers through classroom observations, by the amount their students learn, or by some combination of these and other measures. The evaluations can have high stakes for teachers. Poor evaluations may lead to frozen salaries, mandatory remediation, or dismissal, while exceptional evaluations may be rewarded with salary increases and tenure.

Considering the high stakes for teachers, it is critical to examine the quality of the measures used to evaluate teachers. Recognizing this, Washoe County School District in Reno, Nevada, requested assistance from Regional Educational Laboratory West to examine qualities of the teacher ratings derived from the observation instrument that the district planned to use in evaluating teachers. This study examines the validity of ratings from a classroom observation instrument in a teacher evaluation system adapted (by modifying the wording of some of the rated practices) from the Danielson Framework for Teaching (Danielson, 2007) and proposed for use in Washoe County School District (see box 1 for a summary of the two frameworks). While the study takes place in one school district, the findings may be of interest to districts and states that are using or considering using the Danielson Framework. Previous studies have examined only about half of the Danielson Framework, typically focusing on the ratings of teachers' classroom practices rather than the ratings of teaching dimensions that describe practices occurring outside of classroom instruction. This study is unique in examining all dimensions of the Danielson Framework, including teachers' planning for instruction and carrying out of professional responsibilities.

Box 1. About the Danielson Framework for Teaching and the Washoe County School District–adapted system

The Danielson Framework for Teaching

The Danielson Framework for Teaching is a hierarchical organization of teaching activities that "aims to describe all of teaching … not only … what occurs in the classroom but also … what happens behind the scenes and beyond the classroom walls" (Danielson, 2007, p. 19). The Danielson Framework organizes 22 teaching activities—components of teaching—into four domains:

- Domain 1, Planning and Preparation, identifies teaching components involved in instructional design. These components describe how a teacher organizes content and students for instruction. A teacher's skill in this domain is assessed by examining the teacher's instructional plans and how the teacher describes his or her decisions in creating those plans.
- Domain 2, Classroom Environment, contains the activities that establish a comfortable and respectful classroom environment that cultivates a culture for learning and creates a safe place for risk taking. Observations of classroom interactions and interviews with students provide evidence to assess teacher skill in this domain.
- Domain 3, Instruction, describes teaching activities involved in engaging students in content. Skills in this domain, like those in the Classroom Environment domain, are assessed by observing the teacher's and students' interactions in the classroom. Examination of student work may be used to assess the degree of cognitive challenge expected of students.
- Domain 4, Professional Responsibilities, distinguishes the Danielson Framework from other teacher observation frameworks, according to its developer. This domain identifies activities associated with being a true professional educator that encompass the roles assumed outside of and in addition to those in the classroom with students. The components in this domain include teacher interactions with parents and participation in

(continued)

Box 1. About the Danielson Framework for Teaching and the Washoe County School District–adapted system *(continued)*

> professional communities. Evidence to judge a teacher's skill level in this domain is found in teacher logs and other summaries of their participation in school, district, and professional activities.

See tables A1–A4 in appendix A for a list of the elements in each domain and how the elements in the Danielson Framework correspond to those in the Washoe County School District–adapted system. Each component comprises two to five teaching elements that describe specific features of the component. Rubrics are available for rating teacher skill at the domain and element levels on a four-point scale: unsatisfactory, basic, proficient, or distinguished. Ratings can be recorded and reported at each level (that is, 76 elements, 22 components, four domains) and overall.

Washoe County School District–adapted system

Washoe County School District adapted the Danielson Framework for Teaching for use in the district through the following process:

> WCSD [Washoe County School District] embarked on developing a new teacher evaluation rubric. In establishing this new rubric, over 90 people representing teachers, site administrators, district administrators, local universities, community members, and parents came together over four days in 2011, to work on a final product. The teacher evaluation rubric group used the Charlotte Danielson model as a starting point and refined the document to meet Nevada, as well as WCSD[,] goals and objectives. The final product was vetted by a sub-committee consisting of members who participated in the development of the new rubric. The final product was completed in June of 2011 and moved forward to be piloted at [...] WCSD schools during the 2011/12 school year. (Kendrick, 2012, slide 3)

Like the Danielson Framework for Teaching, the Washoe County School District system's classroom observation instrument has four main ratings associated with the four domains. Each domain comprises 5 or 6 components, for a total of 22 components of classroom teaching (see tables A1–A4 in appendix A for a list of the elements in each domain and how the elements in the Danielson Framework correspond to those in the Washoe County School District–adapted system). While the elements within each component are used to guide the observation, it is the 22 components that observers rate on a four-point scale of effectiveness. The four domain ratings are constructed by averaging the ratings for the components within each domain; a single summative rating is computed by averaging ratings across all 22 components.

For each component, Washoe County School District maintained the Danielson Framework for Learning definitions (Danielson, 2007) of the four points on the rating scale but replaced the labels (unsatisfactory, basic, proficient, and distinguished) with teacher effectiveness labels provided by the Nevada State Department of Education: 1 = ineffective, 2 = minimally effective, 3 = effective, and 4 = highly effective.

What the study examined

This study evaluated the validity of ratings of teachers by their principal using the classroom observation instrument proposed for implementation in Washoe County School District (see box 2 for more on the data used in the study). The district intends to interpret the ratings as indicators of teaching effectiveness both within the district's continuous improvement model and for the state's teacher evaluation system. Within the district's continuous improvement model the district would like to interpret each domain rating as information about a different aspect of teaching, which would enable it to provide teachers

Box 2. Study data

The observation data examined in this study are the ratings, given by principals or assistant principals, for 713 Washoe County School District elementary, middle, and high school teachers who were observed on all 22 components of the classroom observation instrument in the 2012/13 school year. Teachers were observed by their principals on all components if they were probationary teachers in their first three years of teaching or if they were tenured teachers who were scheduled for a mandatory review that occurs every five years. Thus, while observations are not available for the full population of Washoe County School District teachers, those who were observed span a wide range of teaching experience and come from schools from across the district. Data were not available on teacher tenure or assignment, so it is not possible to describe how teachers are distributed by experience level, grade, or school. Data were also not available on the length of time the principal spent with the teacher, the nature of the lessons observed, or the number of times a principal observed a teacher before making a rating.

Washoe County School District provided Regional Educational Laboratory West with teacher-level data files that included all ratings from the classroom observation instrument for each teacher who was observed during the study year. In addition, for the fourth research question about the relationship of ratings from the classroom observation instrument to student outcomes, the district provided a second database that contained student achievement test data that allowed the study team to compute the 2012/13 teacher-level growth scores in reading and math, which are the median of their student growth scores derived by the state from the Nevada Growth Model. The growth scores were available for teachers in grades 4–8 who taught reading or math and who had at least 10 students in their class who had attended Nevada schools in at least one previous year. Both data files coded teachers by a number that was used to merge the two files but could not be used by the study team to identify the teacher.

with information about their areas of strength and their areas of weakness that might be improved with professional development. Within the state's teacher evaluation system, it is planned that observation ratings, along with student achievement ratings, will inform decisions about tenure, retention, and an anticipated pay-for-performance system (Dale Erquiaga, personal communication, August 2014). Thus, there is interest in identifying not only low-scoring teachers who might benefit from professional development, but also high-scoring teachers who might be rewarded for the quality of their teaching.

Washoe County School District would like to interpret each domain rating in its teacher evaluation system as information about a different aspect of teaching

Examining validity requires identifying and assessing the plausibility of the assumptions that underlie the planned use of the results (Cronbach, 1988; Kane, 2006, 2013). This study examined four assumptions underlying Washoe County School District's proposed interpretation of ratings from the classroom observation instrument:

- Ratings from the classroom observation instrument differentiate among teachers.
- Each of the four domain ratings measures a single, cohesive area of teaching practice.
- Each of the four domains is distinct from the others.
- Ratings from the classroom observation instrument indicate teacher effectiveness in promoting student learning.

These four assumptions are examined through analyses that address the study's four research questions, which are described below. A single study cannot validate an

assessment; multiple sources of evidence must be examined. This report thus includes a summary of related findings that other researchers have reported.

What are the distributions of teacher ratings for the 22 components and four domains?

The first research question examines the assumption that ratings from the classroom observation instrument differentiate among teachers. Washoe County School District plans to use the ratings to distinguish low-performing teachers from high-performing teachers in order to identify teachers who would benefit from training and to inform decisions about tenure, retention, and pay for performance. Thus, examining the rating distributions could help the district evaluate whether the instrument distinguishes higher performing teachers from lower performing teachers.

The first research question examines the assumption that ratings from the classroom observation instrument differentiate among teachers; the second research question examines whether there is empirical evidence to support the assumption that each domain rating is a measure of a single, cohesive area of teaching practice

If the ratings differentiate among teachers, there will be variability in the ratings for the domains and components within the domains. The more variability among the ratings, the more the ratings discriminate among teachers. However, if most teachers receive only one of the possible ratings, the instrument would appear to not discriminate between higher and lower performing teachers.

While results from the statistical analyses can provide evidence to support the assumption that ratings discriminate among teachers, they cannot provide conclusive evidence to refute that assumption. If ratings do not vary across teachers, the analyses cannot explain why they do not vary. A lack of discrimination could indicate that the ratings from the classroom observation instrument lack validity in that they are unable to differentiate stronger teaching skills from weaker teaching skills. There are several reasons validity may be limited, including principals rating most teachers highly because they are not comfortable rating their teachers, poor training in the use of the classroom observation instrument, or principals failing to follow the descriptions of the levels of teaching skills when scoring teachers. A lack of discrimination also could indicate a situation in which the ratings are valid but the observed teachers do not differ in teaching skill. In that situation principals may be rating the teachers accurately, but the teachers simply do not differ one from another.

By examining the rating distributions, Washoe County School District can begin to understand how well ratings from the classroom observation instrument differentiate among teachers. Instances in which ratings from the classroom observation instrument do not differentiate among teachers may help generate ideas for future studies as to whether and how principal training or observation implementation might be changed to improve the discrimination of the ratings.

What is the internal consistency of ratings within each domain and across all domains?

The second research question examines whether there is empirical evidence to support the assumption that each domain rating is a measure of a single, cohesive area of teaching practice. For example, to interpret a teacher's rating for domain 3 (Instruction) as an indication that the teacher has a high or low skill level in instruction would mean interpreting all the components in domain 3 as indicators of instructional skill.

The internal consistency of ratings summarizes the relationships among the component ratings. Internal consistency is important because if the component ratings within a domain all measure a common aspect of teaching, they should have strong relationships. That is, teachers who receive a high rating for one component in a domain should receive high ratings for the other components in that domain as well. Thus, when a group of component ratings are internally consistent, it makes sense to summarize the ratings by averaging or adding them because there is reason to believe that they measure a common domain of teaching.

But if the ratings for the components within a domain have low internal consistency, they are not providing similar information about a teacher. Thus, they may not be interpreted as indicators of the same domain. In that situation, one would need to re-evaluate whether a summary score for the domain is meaningful.

If the assumption were supported, the expected pattern of results from this analysis would be high internal consistency within each domain. However, high internal consistency would not necessarily be expected across the rating for all 22 components because the domains are each intended to measure distinct areas of teaching practice (this assumption is further examined through the third research question).

Do the empirical data support the hypothesized grouping of components into domains?

The third research question examines the assumption that each domain encompasses a distinct aspect of teaching practice; the fourth research question addresses the assumption that ratings from the classroom observation instrument indicate teacher effectiveness in promoting student learning

The third research question examines the assumption that each domain encompasses a distinct aspect of teaching practice by analyzing whether the observed relationships among ratings correspond to the theoretical structure of the ratings (that is, a hierarchy of 22 components grouped within four distinct domains). Correspondence would support the interpretation of domain ratings as information about four distinct aspects of teaching practice. It is important to know whether this interpretation of domain ratings is valid because it affects how and to what extent teachers' domain ratings can be interpreted and used. For example, if Washoe County School District wants to plan professional development to improve skills of low-rated teachers, would a teacher's ratings for the four domains identify which aspect of teaching needed improvement? Would it be possible to distinguish teachers who would benefit from training in instructional planning (domain 1) and others who would benefit from training in classroom management and organization (domain 2)? Without evidence that the four domain ratings provide distinct information, this type of interpretation would not be supported.

For the empirical structures to correspond to the theoretical structure of the four domains —that is, for the study's findings to support the assumption that each domain is distinct from the others—the analysis of the relationships among the 22 components should show stronger relationships between components within the same domain than between components from different domains. If this is not the case, the data do not support interpretation of the four domains as representative of separate teaching traits, and Washoe County School District would need to rethink its planned interpretations of the domain ratings.

Depending on the outcomes of the third research question, other interpretations might be made. For example, if the analyses did not support four distinct groups of components —one for each domain—and instead supported one larger group of components, as if all components belonged to one overarching domain, Washoe County School District might

want to consider interpreting the ratings as indicators of general teaching skill. In that scenario, the average rating over all the components, rather than individual domain ratings, could be used to evaluate teachers on general skill.

Do ratings from the classroom observation instrument predict student learning?

The fourth research question addresses the assumption that ratings from the classroom observation instrument indicate teacher effectiveness in promoting student learning. If ratings from the classroom observation instrument are to be interpreted as measures of teacher effectiveness, they need to correlate positively with a measure of student learning.

In Nevada, student learning is measured by a growth model known as the Nevada Growth Model. A student's score in this model describes how much a student achieved during an academic year, relative to all other students in the state who started the year at the same level of achievement as that student.[2] The median growth score for the students in a teacher's class is the teacher's growth score for the year.

The expectation for the analyses of the fourth research question is that if ratings from the classroom observation instrument measure teacher effectiveness in promoting student learning, the ratings will correlate with student growth as measured during the year the students worked with the teacher.

What the study found

This section describes the statistical analyses and findings for each research question and discusses how this study's findings compare with findings from other investigations of the same or similar question.

On each component most teachers were rated as effective or highly effective

With one exception, the distributions of ratings for the 22 components showed that at least 90 percent of teachers were rated as effective or highly effective

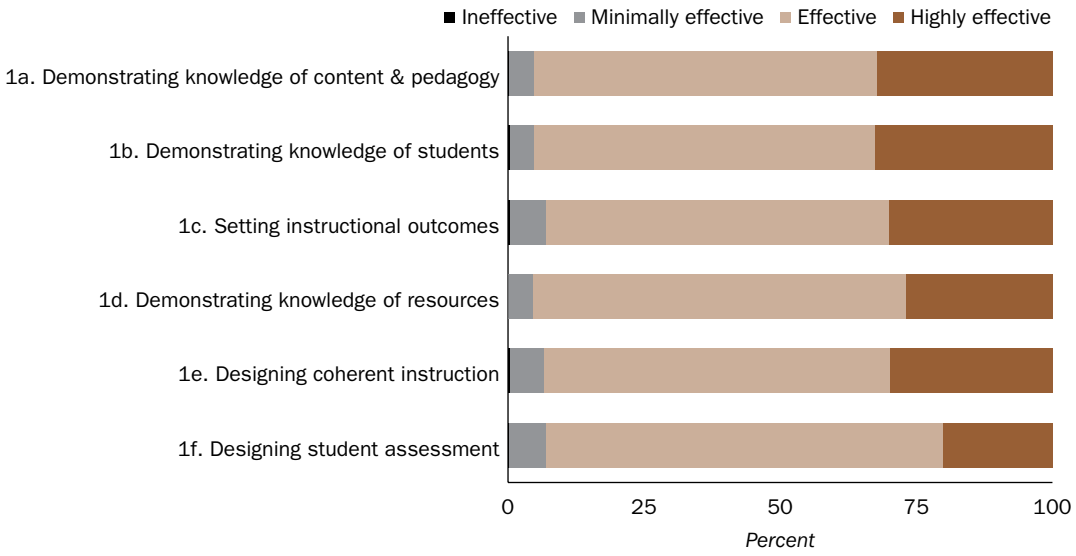
Statistical analyses and findings. With one exception, the distributions of ratings for the 22 components showed that at least 90 percent of teachers were rated as effective or highly effective (figures 1–4). The exception was component 3b, using questioning and discussion techniques, for which 88.4 percent of teachers received one of the two highest ratings (see figure 3).

For all but one component, effective was the most common rating, received by 53.6–72.9 percent of teachers (see figures 1–4). The exception was component 2a, creating an environment of respect and rapport, for which ratings were more evenly split between effective (47.7 percent) and highly effective (49.8 percent; see figure 2).

The average of all 22 components might be used as a single indicator of overall effectiveness, while averages over the ratings within a domain might be used as an indicator of a teacher's effectiveness in that domain. Here the study team considers the distribution of these average ratings.

For overall effectiveness, computed as the average rating over all 22 components in the observations, most teachers (85 percent) were rated at the numerical equivalent of effective (a rating of 3.0) or higher (figure 5). These teachers' average ratings were distributed

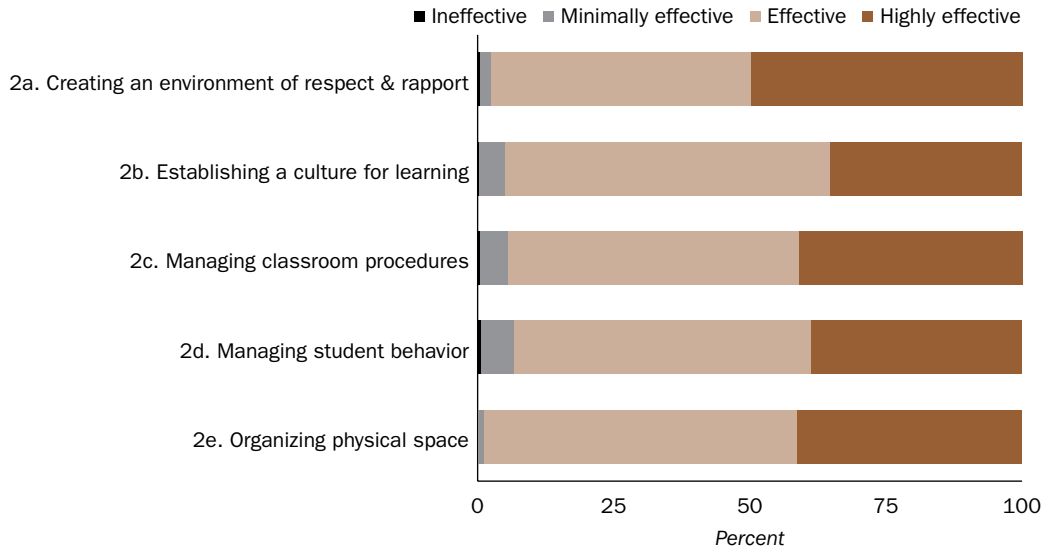
Figure 1. Percentage of teachers by rating for each component in domain 1, Planning and Preparation



Note: *n* = 713. See table A5 in appendix A for percentages and frequencies.

Source: Authors' analysis of data on 2012/13 teacher ratings from Washoe County School District.

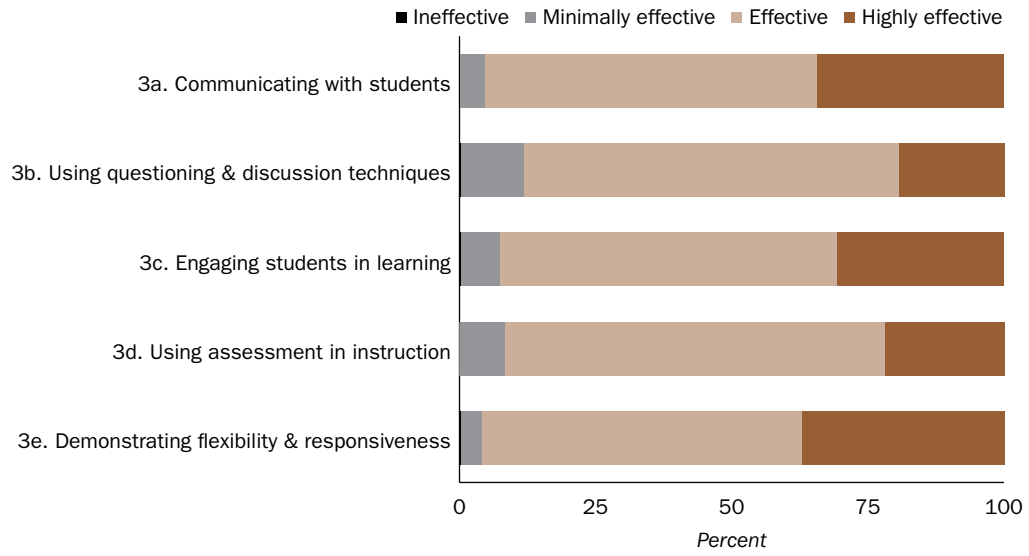
Figure 2. Percentage of teachers by rating for each component in domain 2, Classroom Environment



Note: *n* = 713. See table A5 in appendix A for percentages and frequencies.

Source: Authors' analysis of data on 2012/13 teacher ratings from Washoe County School District.

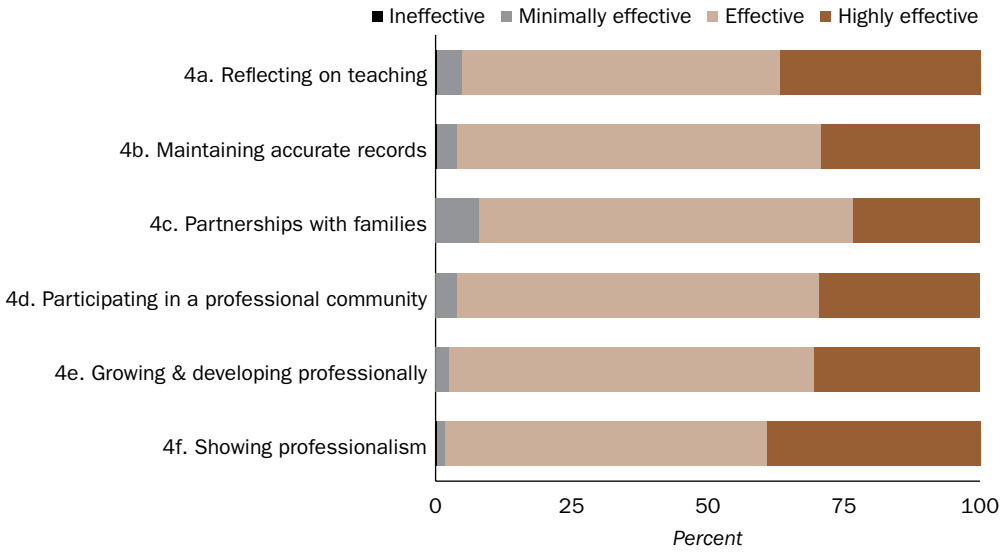
Figure 3. Percentage of teachers by rating for each component in domain 3, Instruction



Note: *n* = 713. See table A5 in appendix A for percentages and frequencies.

Source: Authors' analysis of data on 2012/13 teacher ratings from Washoe County School District.

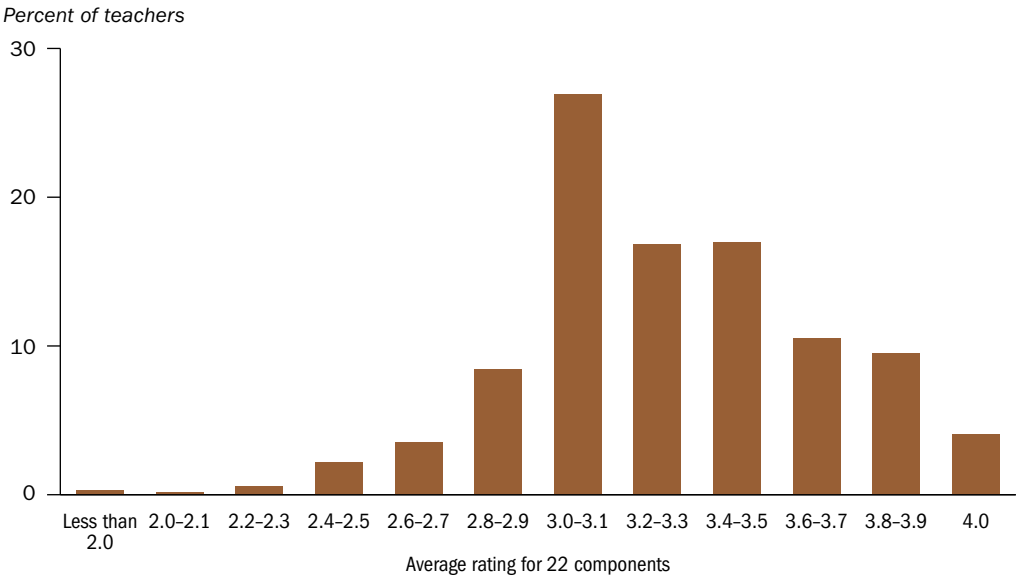
Figure 4. Percentage of teachers by rating for each component in domain 4, Professional Responsibilities



Note: *n* = 713. See table A5 in appendix A for percentages and frequencies.

Source: Authors' analysis of data on 2012/13 teacher ratings from Washoe County School District.

Figure 5. Teachers' averages over all component ratings



Note: *n* = 713. The averages are computed across 22 ratings on a four-point scale, where 1 = ineffective, 2 = minimally effective, 3 = effective, and 4 = highly effective.

Source: Authors' analysis of data on 2012/13 teacher ratings Washoe County School District.

across the range from 3.0 to 4.0, indicating that by aggregating across the component ratings, it is possible to distinguish among the teachers who are rated effective. About 15 percent of teachers received an average rating below the cutoff for effective. Most of those teachers were rated at or above 2.0, the numerical equivalent of minimally effective, and 0.28 percent received an average rating of less than 2.0.

Averages computed within each domain showed distributions similar to that shown in figure 5, in that on each domain at least 80 percent of teachers received an average rating of 3.0 or higher, while less than 1 percent received a rating below 2.0 (see table A6 in appendix A). These distributions, and the one for the the average of all the component ratings in figure 5, convey that principals are discriminating among teachers but are not using the full range of the rating scale. Interpreting the ratings by their labels suggests that principals are discriminating among effective and highly effective teachers but rarely identifying teachers as minimally effective or ineffective. The data cannot explain why this is. Washoe County School District officials may want to consider how well these results compare with their expectations and, if they anticipated different results, explore with principals why the ratings resulted in the distributions that were found.

Interpreting the ratings by their labels suggests that principals are discriminating among effective and highly effective teachers but rarely identifying teachers as minimally effective or ineffective

How the findings compare to findings from other research studies. Recent studies have examined the distributions of ratings from observations based on the Danielson Framework for Teaching. Like this study, the other studies found that raters tended to use only a few points on the four-point scale. However, in comparing the results of this study and the others, there appear to be differences between the way Washoe County School District principals used the scale and the way raters in the other studies used the scale. These differences, detailed below, suggest that Washoe County School District may want to examine further whether their principals are rating based on the labels given to ratings or on the expanded descriptions of teaching behavior at each of the points.

Sartain, Stoelinga, and Brown (2011) reported on ratings given to the same Chicago Public Schools teachers by the teachers' principals and by external observers who did not know the teachers. The external observers in Chicago used primarily the two middle points on the scale (rating levels 2 and 3—labeled in the Danielson Framework as basic and proficient, respectively), rating at least 90 percent of the teachers at those two points. Component ratings by the principals showed a slightly different pattern. The principals frequently used the second highest rating, level 3 (at least 49 percent of teachers were rated at level 3 by principals), but split their ratings of the remaining teachers between levels 2 and 4. The same pattern was found in a study of principal ratings of approximately 900 Pittsburgh teachers conducted by Chaplin, Gill, Thompkins, and Miller (2014). In that study at least 57 percent of teachers were rated at level 3 on each component, and principals split the ratings of the remaining teachers between levels 2 and 4.[3]

Kane and Staiger (2012) reported rating distributions for 8 of the 10 components of domains 2 and 3 for approximately 1,300 teachers from six districts across the country in observations conducted by trained observers who did not know the teachers. Like the Chicago external observers, the observers in Kane and Staiger's study used primarily the middle two points of the scale (levels 2 and 3)—rating at least 90 percent of the teachers at these points.

In future studies Washoe County School District may want to examine why its principals tended to use higher rating points than raters in other studies

Washoe County School District principals, Chicago external observers, and the external observers in Kane and Staiger (2012) tended to use only two of the four rating points when evaluating teachers. However, the studies differed in that the Washoe County School District principals rated teachers primarily on the highest two points, while the external observers in Chicago and the observers in Kane and Staiger's study used primarily the middle two points, rarely giving teachers the highest rating.

In future studies Washoe County School District may want to examine why its principals tended to use higher rating points than raters in the other studies. One difference in the studies is that the external observers in Chicago and the raters in Kane and Staiger (2012) did not know the teachers they were rating. The fact that principals work with the teachers on a daily basis may influence how they use the rating scale. There is some evidence to support this hypothesis in the Chicago study, in which principals were found to give higher ratings than external observers to teachers who had received high ratings from evaluations in previous years. Thus, the Chicago principals may have taken into account prior knowledge of the teachers that was not available to the external raters.[4]

Another hypothesis to examine is whether the adaption that Washoe County School District made to the labels of the rating scale made the principals more inclined to use the higher points on the scale. Whereas the original scale, used in the two other studies, labeled scores 2 and 3 as basic and proficient, respectively, Washoe County School District relabeled those scores as minimally effective and effective. It may be worth examining whether Washoe County School District labels are interpreted by principals in the same way as the original labels of the four-point rating scale.

The component ratings are consistent in the information they provide about teachers

Statistical analyses and findings. The analyses summarized the internal consistency of ratings by a measure known as Cronbach's alpha (α), an index that ranges from 0 to 1, with a higher number indicating greater internal consistency.

Within each domain, principals were consistent in their scoring of teachers (table 1). Teachers who received a high rating for one component tended to receive a high rating for the other components in the domain as well; those who received a low rating for one component tended to also receive a low rating for the other components. The fact that the internal consistency was higher when all components were pooled across all domains suggests that components from different domains are also highly related, such that the ratings for the four domains may not be providing information about different aspects of teaching.[5] Instead, the ratings may reflect a single dimension that could be best captured by an average rating taken over all components—a possibility tested more directly in the third research question.

How the findings compare to findings from other research studies. Two recent studies also examined the internal consistency of ratings from the Danielson Framework for Teaching. Chaplin et al. (2014) evaluated the internal consistency of ratings for a subset of 12 components that were selected by their study district because the district educators believed they were especially influential in promoting student learning. They found an internal consistency of .87, similar to those in table 1.

Teachers who received a high rating for one component tended to receive a high rating for the other components in the domain as well; those who received a low rating for one component tended to also receive a low rating for the other components

Milanowski's (2011) study of Washoe County School District teacher observations based on the Danielson Framework for Teaching examined internal consistency for two domains: domain 1 (Planning and Preparation) and domain 2 (Classroom Environment). The study was designed primarily to examine agreement of ratings of teachers by their peers and by administrators, conducted on different days. To examine internal consistency of the domain ratings, Milanowski first averaged the peer and administrator ratings for each teacher. He then computed the internal consistency of those average component ratings. The internal consistency was .78 for domain 1 and .82 for domain 2. Milanowski's findings are slightly lower than the findings in table 1 and those in Chaplin et al. (2014). This could be due to the fact that his ratings for each teacher are based on perceptions of two raters, from different roles (whose agreement may vary from one component to another), while the other findings are based on ratings from the principal only.

The empirical data do not support grouping ratings into four domains

Statistical analyses and findings. Confirmatory factor analysis was used to test specific hypotheses about the relationships among the component ratings. This method requires that the researcher specify a theory about how the components should relate to one

Table 1. Internal consistency for 2012/13 ratings by domain

Domain	Number of components rated	Internal consistency (Cronbach's alpha)
1. Planning and Preparation	6	.87
2. Classroom Environment	5	.85
3. Instruction	5	.82
4. Professional Responsibilities	6	.83
Total over all domains	22	.95

Note: n = 713.

Source: Authors' analysis of the internal consistency of 2012/13 teacher ratings from Washoe County School District.

another (for example, if the four domains measure different aspects of teaching, the component ratings would be expected to group into the four domains). The method can then test several different specifications (that is, specifications that differ in the number of item groupings or in the items in the groups) and determine which hypotheses best match the available data. The hypotheses tested, as well as the assessment of their match to the data, are in appendix B.

The results did not support the interpretation of the four domains of the Danielson Framework for Teaching as four distinct aspects of teaching. Although the hypothesis that components would group into the four domains of the Danielson Framework matched the data reasonably well, ratings for the four groupings were so highly correlated as to suggest that each group measured a common feature of teaching. Thus, all components appeared to belong to a single group, and only one overall rating, derived from all the components, would be needed to summarize a teacher's evaluation.

Ratings for the four groupings were so highly correlated as to suggest that each group measured a common feature of teaching. Thus, all components appeared to belong to a single group, and only one overall rating, derived from all the components, would be needed to summarize a teacher's evaluation

There are two interpretations of this finding of a single grouping of the components, but their implications for practice are the same. One interpretation is that the four domain ratings do not provide information about different aspects of teaching. The other is that they do provide information about different aspects of teaching but that those aspects are so highly correlated that knowing about one aspect provides information about the others. In either case the analysis does not support interpreting the four domain scores as measurements of distinct aspects of teaching; instead, the analysis supports using a single rating, such as the average over all components of the system to summarize teacher effectiveness.

How the findings compare to findings from other research studies. The study team was unable to identify other research that specifically tested the hypothesis of a hierarchical organization of teaching components within domains, as this study has done.

Teachers' scores on the classroom observation instrument are related to measures of their students' learning

Statistical analyses and findings. The study team used Pearson correlations to summarize the relationship between teacher ratings from the classroom observation instrument and growth, as measured by the teacher-level score from the Nevada Growth Model. If teacher ratings from the classroom observation instrument measure teacher effectiveness in promoting learning, the ratings should correlate with student growth as measured during the year the students worked with the teacher. The analysis is based on subsamples of the 713 teachers. Each correlation includes only teachers who taught in grades 4–8 (where annual student achievement growth scores can be computed) and who had at least 10 students in their class.

The correlations of the score on the classroom observation instrument with student growth indicated a positive and significant relationship existed in all but one case (table 2). The exception was the correlation between teacher ratings for domain 4 (Professional Responsibilities) and growth in reading. These findings support the interpretation of the scores as measures of teacher effectiveness.

How the findings compare to findings from other research studies. Previous studies have examined how observation ratings of teachers on the Danielson Framework for Teaching,

Table 2. Pearson correlation coefficients of domain ratings for 2012/13 with student growth, by subject

Domain	Growth	
	Reading (n = 113)	Math (n = 118)
1. Planning and Preparation	.23* (.04, .40)	.39** (.23, .53)
2. Classroom Environment	.38** (.21, .52)	.45** (.30, .59)
3. Instruction	.28** (.10, .44)	.48** (.32, .60)
4. Professional Responsibilities	.18 (–.00, .36)	.37** (.20, .52)
Total over all domains	.29** (.11, .45)	.46** (.30, .59)

* Significant at $p < .05$. ** Significant at $p < .01$.

Note: Numbers in parentheses are the bounds of the 95 percent confidence interval for the correlation coefficient. Teacher scores, derived from the Nevada Growth Model, are the median growth score for students the teacher taught. Only teachers who taught reading or math in grades 4–8 and had at least 10 students with growth scores were included in the computation of the correlation coefficient.

Source: Authors' analysis of data on 2012/13 teacher ratings from Washoe County School District.

The correlations of the score on the classroom observation instrument with student growth indicated a positive and significant relationship existed in all but one case

or an adaption of that framework, relate to student learning as estimated by value-added measures of student growth. The study team found no studies that related the ratings under the Danielson Framework (or an adaptation thereof) to growth as measured by a student growth percentile model, such as the Nevada Growth Model used in Washoe County School District.

Statistically significant relationships in the expected direction were found in most of the studies reviewed. Researchers used different methodologies to summarize the relationships. As in this study, some examined correlations of observation ratings and student learning. Others contrasted the average learning in classes of teachers who had higher and lower ratings. The research studies, their findings, and methodologies are summarized below so that readers can consider whether these methods should be used in other analyses or future studies in Washoe County School District.

An earlier study of Washoe County School District elementary school teachers reported correlations between the average rating received by a teacher on the observation components and that teacher's value-added scores in reading and math computed by the researcher (Milanowski, 2011). Correlations ranged from .19 to .24 depending on the year (2001–03) and subject (reading or math); all correlations differed from zero at a significance level of $p < .05$.

Chaplin et al. (2014) studied principal ratings of 358 elementary and middle school teachers and examined the relationship of principal ratings and teachers' value-added estimates. In reading, the correlation between a teacher's average ratings for all 22 components, and the teacher's value-added estimate of achievement was .29 (statistically significant at $p < .05$). In math, the correlation was not significantly different from zero.

Kane, Taylor, Tyler, and Wooten (2011) examined the ratings teachers received on the components of domains 2 and 3 in a study conducted in Cincinnati Public Schools. For each component they compared the ratings received by teachers who were in the lowest and highest quartiles in the distribution of value-added measures in reading and math. They then compared the teachers in the second quartile with those in the highest quartile. All

comparisons but one differed from zero at a significance level of $p < .05$, and the difference was in the expected direction. That is, teachers who had higher value-added scores were teachers who, on average, received higher ratings for domains 2 and 3 of the observation.

Sartain et al.'s (2011) study of the Danielson Framework for Teaching in Chicago Public Schools examined ratings of 417 teachers in reading and 340 teachers in math. In each subject, for each component in domains 2 and 3 teachers were grouped according to the rating the teacher received on the component (unsatisfactory, basic, proficient, or distinguished). To determine whether there were differences, on average, in value-added measures across the four teacher groups, a one-way analysis of variance was conducted. In reading, the group differences were statistically significant at $p < .05$ for all but one of the 10 components. Additionally, for all but one of the significant components, the expected linear trend was found for the group means: the lowest mean was found for the group that received an unsatisfactory rating, the next lowest was for the group that received a basic rating, the next lowest was for the group that received a proficient rating, and the highest was for the group that received a distinguished rating. In math, differences across the four groups were statistically significant at $p < .05$ for each component, and each component showed the expected linear trend.

Previous studies have examined how observation ratings of teachers on the Danielson Framework for Teaching, or an adaption of that framework, relate to student learning as estimated by value-added measures of student growth. Statistically significant relationships in the expected direction were found in most of the studies reviewed

Kane and Staiger (2012) examined the relationship of value-added measures of teachers' contributions to student learning with ratings of teachers on the components in domains 2 and 3 of the Danielson Framework in the Measures of Effective Teaching Project. They controlled for classroom factors (other than teacher effectiveness) that might affect both the observers' ratings and student value-added scores and thus create a spurious correlation. Kane and Staiger compared the observation ratings taken in one classroom to the value-added score of the teacher as measured in a different classroom. In one analysis Kane and Staiger correlated the teacher's average rating across domains 2 and 3 with the average gain made by the teacher's students the prior year. The correlation was .07 for English language arts and .13 for math. A second analysis examined the relationship of the rating from an observation taken in one section of a course with the gain of students in a second section of the course taught in the same year. The correlation was .05 for English language arts and .13 for math. None of the correlations was significantly different from zero.

Although Kane and Staiger did not find relationships between teachers' observation ratings and value-added scores across the full sample of teachers, they did find that the observation ratings distinguished teachers at the extremes of the distribution of value-added scores. Kane and Staiger contrasted the mean observation rating of teachers whose students were in the upper and lower quartiles of the distribution of achievement gain. For both subjects (English language arts and math) and both types of classes (prior year and same year but different section), differences were statistically significant at $p < .05$, ranging from .02 to .07 standard deviation unit.

Implications of the study findings

Findings of this study, especially as they are considered along with other researchers' findings, have three main implications for interpretation, use, and future research into the ratings derived from the Danielson Framework for Teaching.

First, the study found that principals discriminated among teachers but did not use the full range of the rating scale. Specifically, principals discriminated among those they thought to be effective and highly effective, as labeled by the Washoe County School District–adapted system; they rarely identified teachers as minimally effective or ineffective. Other research also found that some raters failed to use the full range of the rating scale, but unlike Washoe County School District rates, they tended to use a different portion of the scale and to discriminate between ratings of basic and proficient (as labeled in the original Danielson Framework). Given these findings, Washoe County School District may want to consider how well the results compare with their expectations and, if they anticipated different results (for example, that more teachers would be rated as ineffective), explore with principals why the ratings resulted in the distributions that were found. It also may be that Washoe County School District's re-labeling of the scale influenced how principals interpreted the ratings.

Washoe County School District may want to consider how well the results compare with their expectations and, if they anticipated different results, explore with principals why the ratings resulted in the distributions that were found

Second, the study findings have implications for the interpretation of average ratings that are created from the component ratings. The findings support the use of a single summative rating for a teacher derived by averaging ratings for the 22 components. They do not support using domain or component ratings to evaluate teachers' skills because there was little evidence that the ratings measure distinct aspects of teaching as hypothesized in the construction of the Washoe County School District–adapted system. Instead, they all are highly related and function as if they are measuring something in common to all. Why this happens cannot be determined by the data in this study. It may be that the hypothesized domains are not distinct but instead contain similar or overlapping teaching skills. Or it may be that they are distinct but that principals are not applying the rubrics of the system with fidelity. In either case the domain scores do not appear to provide distinct information about different teacher practices or skills that would be useful to identify teachers' strengths and weaknesses and guide future professional development, as Washoe County School District had hoped.

Finally, the information the domain ratings provide, individually and in total, predicts the growth scores of the teachers' students. Assuming the growth scores are sound measures of student learning, this finding offers some evidence that the observation ratings provide information about a teacher's skill in promoting learning and can add to Washoe County School District's confidence in interpreting a teacher's rating as a measure of effectiveness. Still, other factors could cause this positive relationship between teacher rating from the classroom observation instrument and students' growth score, and future research into the relationship could be conducted to examine it further.

Limitations of the study

This study has several limitations. First, the sample derives from a subset of teachers (only probationary teachers and tenured teachers in an evaluation year) from a single school district, which limits the generalizability of the findings. The analysis for the third research question concerning the theoretical grouping of components into four dimensions may be limited by the need to combine the two lowest ratings on the four-point rating scale for components because on each component too few teachers received the lowest rating. In addition, the analysis for the fourth research question included only teachers in the sample for whom growth scores could be computed—that is, those in grades 4–8 who taught reading or math and who had at least 10 students in their class who had attended

Nevada schools in at least one previous year. Another limitation is that Washoe County School District adapted the Danielson Framework for Teaching, so comparisons to the standard version should take this into consideration.

While not limiting the study's ability to address the research questions, the available data could not answer some questions that would have been informative to address. Data availability precluded describing how teacher performance was distributed by experience level, grade, or school and analyses of rater reliability or the fidelity or quality of the observation process. And since only one year of observation data was available, the study team could not assess the extent to which teachers' observed performance in one year predicted their subsequent performance.

Appendix A. Supporting tables

This appendix provides a table for each domain of the Danielson Framework for Teaching that compares the original Danielson Framework to the Washoe County School District–adapted system (tables A1–A4) as well as tables that provide the data used to create figures 1–5 in the main text (tables A5 and A6).

Table A1. Comparison of domain 1, Planning and Preparation, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system

Danielson Framework for Teaching	Washoe County School District adapted system
Component 1a: Demonstrating knowledge of content and pedagogy	
Knowledge of content and the structure of the discipline	Knowledge of discipline and Common Core State Standards*
Knowledge of prerequisite relationships	No change
Knowledge of content-related pedagogy	No change
Component 1b: Demonstrating knowledge of students	
Knowledge of child and adolescent development	No change
Knowledge of the learning process	No change
Knowledge of students' skills, knowledge, and language proficiency	No change
Knowledge of students' interests and cultural heritage	No change
Knowledge of students' special needs	No change
Component 1c: Setting instructional outcomes	
Value, sequence, and alignment	No change
Clarity	No change
Balance	Integration*
Suitability for diverse learners	No change
	Align outcomes with current standards (new)*
Component 1d: Demonstrating knowledge of resources	
Resources for classroom use	No change
Resources to extend content knowledge and pedagogy	No change
Resources for students	No change
Component 1e: Designing coherent instruction	
Learning activities	No change
Instructional materials and resources	No change
Instructional groups	No change
Lesson and unit structure	No change
Component 1f: Designing student assessment	
Congruence with instructional outcomes	No change
Criteria and standards	No change
Design of formative and summative assessment	No change
Use for planning	Use of assessment in ongoing planning*

* indicates a modification by Washoe County School District.

Note: The Danielson Framework for Teaching was developed by Danielson (2007).

Source: Authors' comparison of the original Danielson Framework and Washoe County School District's adapted framework.

Table A2. Comparison of domain 2, Classroom Environment, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system

Danielson Framework for Teaching	Washoe County School District adapted system
Component 2a: Creating an environment of respect and rapport	
Teacher interaction with students	Teacher/student interaction: Positive regard*
Student interactions with other students	No change
Component 2b: Establishing a culture for learning	
Importance of the content	No change
Expectations for learning and achievement	No change
Student pride in work	No change
Component 2c: Managing classroom procedures	
Management of instructional groups	No change
Management of transitions	No change
Management of materials and supplies	No change
Performance of non-instructional duties	No change
Supervision of volunteers and paraprofessionals	No change
Component 2d: Managing student behavior	
Expectations	No change
Monitoring of student behavior	No change
Response to student misbehavior	No change
Component 2e: Organizing physical space	
Safety and accessibility	No change
Arrangement of furniture and use of physical resources	No change
	Resource-rich environment (new)*

* indicates a modification by Washoe County School District.

Note: The Danielson Framework for Teaching was developed by Danielson (2007).

Source: Authors' comparison of the original Danielson Framework and Washoe County School District's adapted framework.

Table A3. Comparison of domain 3, Instruction, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system

Danielson s Framework for Teaching	Washoe County School District adapted system
Component 3a: Communicating with students	
Expectations for learning	No change
Directions and procedures	Directions, procedures, and explanation of content*
Explanations of content	(Merged into Directions, procedures, and explanation of content)*
Use of oral and written language	(Dropped)*
Component 3b: Using questioning and discussion techniques	
Quality of questions	No change
Discussion techniques	Discussion techniques/student participation*
Student participation	(Merged into Discussion techniques/student participation)*
Component 3c: Engaging students in learning	
Activities and assignments	No change
Grouping of students	No change
Instructional materials and resources	No change
Structure and pacing	No change
	Instructional strategies (new)*
Component 3d: Using assessment in instruction	
Assessment criteria	No change
Monitoring of student learning	No change
Feedback to students	No change
Student self-assessment and monitoring of progress	No change
Component 3e: Demonstrating flexibility and responsiveness	
Lesson adjustment	No change
Response to students	No change
Persistence	No change

* indicates a modification by Washoe County School District.

Note: The Danielson Framework for Teaching was developed by Danielson (2007).

Source: Authors' comparison of the original Danielson Framework and Washoe County School District's adapted framework.

Table A4. Comparison of domain 4, Professional Responsibilities, in the Danielson Framework for Teaching and in the Washoe County School District–adapted system

Danielson Framework for Teaching	Washoe County School District–adapted system
Component 4a: Reflecting on teaching	
Accuracy	No change
Use in future teaching	No change
Component 4b: Maintaining accurate records	
Student completion of assignments	No change
Student progress in learning	No change
Non-instructional records	No change
Component 4c: Communicating with families	Component 4c: Partnership with families
Information about the instructional program	Helping families to navigate the educational system*
Information about individual students	Sharing information about the instructional program and helping families*
Engagement of families in the instructional program	Building partnerships and outreach with families*
	Understanding cultural differences (new)*
Component 4d: Participating in a professional community	
Relationships with colleagues	No change
Involvement in a culture of professional inquiry	Involvement in a culture of professional collaboration*
Service to the school	No change
Participation in school and district projects	No change
Component 4e: Growing and developing professionally	
Enhancement of content knowledge and pedagogical skill	No change
Receptivity to feedback from colleagues	Receptivity to feedback*
Service to the profession	No change
Component 4f: Showing professionalism	
Integrity and ethical conduct	No change
Service to students	Address student needs*
Advocacy	(Dropped)*
Decision making	No change
Compliance with school and district regulations	No change

* indicates a modification by Washoe County School District.

Note: The Danielson Framework for Teaching was developed by Danielson (2007).

Source: Authors' comparison of the original Danielson Framework and Washoe County School District's adapted framework.

Table A5. Distribution of 713 teacher ratings, by domain and component

	Ineffective		Minimally effective		Effective		Highly effective	
Component	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Domain 1								
1a	1	0.14	33	4.63	449	62.97	230	32.26
1b	2	0.38	31	4.35	447	62.69	233	32.68
1c	2	0.28	47	6.59	450	63.11	214	30.01
1d	0	0.00	32	4.49	489	68.58	192	26.93
1e	2	0.28	45	6.31	453	63.53	213	29.87
1f	1	0.14	49	6.87	520	72.93	143	20.06
Domain 2								
2a	3	0.42	15	2.10	340	47.69	355	49.79
2b	3	0.42	33	4.63	425	59.61	252	35.34
2c	3	1.42	36	5.05	382	53.58	292	40.95
2d	5	0.70	43	6.03	388	54.42	277	38.85
2e	1	0.14	8	1.12	409	57.36	295	41.37
Domain 3								
3a	1	0.14	32	4.49	435	61.01	245	34.36
3b	2	0.28	81	11.36	492	69.00	138	19.35
3c	2	0.28	50	7.01	442	61.99	219	30.72
3d	0	0.00	59	8.27	498	69.85	156	21.88
3e	2	0.28	26	3.65	420	58.91	265	37.17
Domain 4								
4a	2	0.28	32	4.49	416	58.35	263	36.89
4b	2	0.28	25	3.51	478	67.04	208	29.17
4c	0	0.00	57	7.99	489	65.58	167	23.42
4d	0	0.00	28	3.93	474	66.45	211	29.59
4e	0	0.00	18	2.52	477	66.90	218	30.58
4f	2	0.28	9	1.26	422	59.19	280	39.27

Note: Percentages may not sum to 100 because of rounding. See tables A1–A4 for component names.

Source: Authors' descriptive analysis of the four-point scale rating, based on 2012/13 data from Washoe County School District.

Table A6. Distribution of 713 teachers' average ratings for four domains

	Teaching domains			
Average rating	1. Planning and Preparation	2. Classroom Environment	3. Instruction	4. Professional Responsibilities
Less than 2.0	0.28	0.56	0.28	0.28
2.0–2.1	1.12	0.28	1.26	0.14
2.2–2.3	2.10	0.98	0.84	0.98
2.4–2.5	0.84	0.98	2.52	0.70
2.6–2.7	3.09	1.12	3.37	2.10
2.8–2.9	7.01	7.43	8.70	8.84
3.0–3.1	31.00	24.82	28.19	26.93
3.2–3.3	24.40	10.24	14.13	13.34
3.4–3.5	8.42	10.80	13.46	10.10
3.6–3.7	7.71	14.17	13.32	9.26
3.8–3.9	6.17	12.34	8.42	8.56
4.0	7.85	16.27	5.33	6.45

Note: Percentages may not sum to 100 because of rounding.

Source: Authors' analysis of data on 2012/13 teacher ratings from Washoe County School District.

Appendix B. Confirmatory factor analyses: Methods and findings

The analyses examined the fit of three competing models that varied based on how the relationships between the components are modeled. The first was a four-factor (group) model in which the components were grouped according to their domain in the Danielson Framework for Teaching (Danielson, 2007). The second was a two-factor model. The first factor included all components from domains 2 and 3, which are ratings of teaching practices observed in classrooms, and the second factor included the components from domains 1 and 4, which are teaching practices that take place outside the classroom and are not based on observation. This model was tested to determine whether the source of information on which ratings were based might be key in determining the rating. The last model was a single-factor model that specified all components formed one group. Because for most components there were fewer than three teachers who received the lowest rating (see table A5 in appendix A), the analyses collapsed the four-category rating scale into three categories by merging the lowest two categories.

The analyses were run with MPLUS software using the weighted least square mean and variance estimator, the default for categorical indicators. To evaluate the models, four fit statistics were examined (table B1). The most common index of fit is the chi-square index (χ^2). Although it is widely used, it is also known for its sensitivity to sample size. Therefore the study team used three additional indices that are relatively independent of sample size to examine model fit: root mean square error of approximation, Tucker-Lewis index, and comparative fit index. In all cases the fit statistics agree that the tested models are reasonable fits to the data.

In addition to the tests of fit of the individual models summarized in table B1, the study team tested pairs of models using DIFFTEST, an MPlus command, which provides a corrected chi-square difference test for two nested models (because when the weighted least square mean and variance estimator is used, the difference in chi-square values as reported in table B1 is not itself distributed as chi-square and thus cannot be compared meaningfully between the models). The corrected values followed the pattern of the results shown in table B1. That is, when compared to the single-factor model, the two-factor and the four-factor models provided better fit to the data, and the four-factor model provided better

Table B1. Confirmatory factor analysis of one- and two-factor models (*n* = 713)

Model description	Chi-square index	Degrees of freedom	Root mean square error of approximation	Tucker-Lewis index	Comparative fit index
One factor	648.71	209	.054	.982	.984
Two factors	527.53	208	.046	.987	.988
Four factors	491.73	203	.045	.988	.989

Note: The critical value for chi-square at $p <$ .05 with 209 degrees of freedom is 243.7; with 208 degrees of freedom it is 242.6; and with 203 degrees of freedom it is 237.4. A root mean square error of approximation of less than or equal to .05 is considered a good fit and less than or equal to .08 is considered a reasonable fit. Anything greater than .10 is considered a poor fit. The Tucker-Lewis index and comparative fit index range from 0 to 1, and a value of .90 or greater indicates a close or adequate fit.

Source: Authors' confirmatory factor analysis of 2012/13 data from Washoe County School District.

fit than the two-factor model. Each comparison was statistically significant at $p < .05$, indicating that the difference in fit was greater than expected by chance alone.

However, the factors in the two-factor and four-factor solutions were so highly correlated as to be nearly redundant. A correlation of 1.0 would indicate the two factors are redundant, providing identical information about teachers. In practice, since a correlation of 1.0 is extreme, researchers often accept correlations above .90 as evidence that two factors are measuring the same domain (Brown, 2006). In the two-factor model the correlation between factors was .93. In the four-factor model correlations for all but one pair of factors was greater than .90 (table B2). Thus, although the components may form different groups, the groups appear to be measuring one aspect of teaching rather than different domains.

Table B2. Intercorrelations of factor scores from a four-factor solution of observation ratings (*n* = 713)

Factors	2	3	4
1	.90	1.00	.93
2		.98	.88
3			.93
4			

Source: Authors' analysis of 2012/13 data from Washoe County School District.

Notes

1. The Danielson Framework for Teaching divides teaching activities into 22 components in four domains of responsibility: Planning and Preparation (6 components), Classroom Environment (5 components), Instruction (5 components), and Professional Responsibilities (6 components). For more information, see http://danielsongroup.org/framework/.
2. The Nevada Growth Model is a student growth percentile model developed by Betebenner (2011). Castellano and Ho (2013) describe the model along with other popular models used to assess teacher effectiveness, such as value-added models.
3. Lazarev, Newman, and Sharp (2014) also examined the distributions of principal component ratings for 297 teachers and found that 57–83 percent of teachers were rated at level 3 on each component. However, because they do not provide the percentages of teachers rated at other levels of the scale, it is not possible to determine whether the same pattern appeared in their findings.
4. Sartain et al. (2011) do not report whether previous evaluations were made by the same principal who was making the current rating. Principal's knowledge about the teacher, prior to the current year evaluation, could come from their previous observations of the teacher or from review of administrative records of previous observations.
5. Cronbach's alpha, like other measures of internal consistency, is influenced not only by the intercorrelation among the ratings, but also by the number of items rated. Other things being equal, if the number of components that are rated increases—without adding components that measure a different, unrelated trait—internal consistency would increase. But if components are added that measure a different trait, the internal consistency would not be expected to increase.

References

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. Dover, NH: National Center for the Improvement of Educational Assessment.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Castellano, K. E., & Ho, A. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in 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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement, 4th edition* (pp. 18–64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. http://eric.ed.gov/?id=EJ996447

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. Retrieved January 21, 2013, from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using achievement data. *Journal of Human Resources*, *46*(3), 587–613.

Kendrick, A. (2012). *Teacher professional growth system*. Presented at the A&S Meeting. Reno, NV: Washoe County School District. Retrieved April 19, 2013, from http://washoecountyschools.org/docs/Teacher_professional_growth_system_powerpoint_8.8.12.pptx.

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.

Milanowski, A. (2011, April). *Validity research on teacher evaluation systems based on the Framework for Teaching.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011, November). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation* (Research Report). Chicago, IL: Consortium on Chicago School Research at the University of Chicago. Retrieved April 19, 2013, from http://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

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