

Introduction

In 2010, Senate Bill 10-191 was passed changing the way teachers are supported and evaluated in Colorado with the ultimate goal of ensuring college and career readiness for all students. To support school districts in implementing the new evaluation requirements, the Colorado Department of Education (CDE) developed a model system as an option for districts to use for teacher evaluations. The Colorado State Model Evaluation System for teachers was piloted in 26 school districts of varying size and location during the 2012-2013 school year (25 of those districts submitted evaluation ratings to CDE). Data from approximately 1,900 teachers in 164 schools that participated in the 2012-2013 pilot of the teacher model evaluation system are presented in this report.

All findings should be considered preliminary for the following reasons:

1. 2012-2013 was the first year educators had any experience with the teacher model evaluation system. Therefore, educators and their evaluators were learning the system while they were using it. Unfamiliarity can affect ratings in different ways.
2. The conversations that result from ratings on the professional practice rubric can be challenging, especially for evaluators who are unfamiliar or uncomfortable with identifying and discussing areas that need improvement. Evaluators who find these conversations challenging, or have limited prior experience facilitating constructive conversations, may assign more positive ratings.
3. This report does not incorporate any specific measures of implementation fidelity. It is likely that some educators received more training on the evaluation system than educators in other districts or other schools. It is also likely that evaluators implemented the evaluation process with varying levels of fidelity. Training and implementation fidelity can affect ratings as well.
4. There is always room for improvement, including improvements to the professional practice rubric. CDE is conducting ongoing analyses of quantitative and qualitative data to ensure that the rubric is not biased based on district, school, teacher, or student characteristics. 2012-2013 was the first year CDE was able to collect data and conduct such analyses.

Summary of Key Findings

- The distributions of teacher ratings across elements and Quality Standards indicate that the professional practice rubric captures multiple aspects of teaching as well as differences in teacher practice.
- The variability in the distribution of ratings suggests that principals (or other teacher evaluators) are able to differentiate between teachers and assign ratings in a meaningful way (e.g., not just assigning the same ratings within or across teachers).
- Colorado teachers get the highest ratings on Standards 2 (Establish Environment) and 5 (Demonstrate Leadership) and the lowest ratings on Standard 3 (Facilitate Learning).

- The majority of teachers improved their ratings throughout the school year.
- Teacher ratings vary based on the district, school level, subject taught, probationary status, experience, and demographic characteristics.

Review of the Colorado Teacher Quality Standards

Before reviewing findings from the first year of the pilot, it is important to consider what exactly comprises the Colorado Teacher Quality Standards. The Quality Standards are the foundation of the professional practice rubric. Note that the rubric measures Standards 1 through 5 (summarized in Figure 1). Standard 6, which captures teacher responsibility for student academic growth, is not assessed by the professional practice rubric and was not piloted during the 2012-2013 school year.

Figure 1. Colorado Teacher Quality Standards and corresponding elements

Quality Standard I: Teachers demonstrate mastery of and pedagogical expertise in the content they teach.

Element a: Instruction that is aligned with the standards and the individual needs of their students.

Element b: Knowledge of student literacy development in reading, writing, speaking and listening.

Element c: Knowledge of mathematics development.

Element d: Knowledge of the content, central concepts, tools of inquiry, instructional practices and specialized character of the disciplines being taught.

Element e: Lessons that reflect the interconnectedness of content areas/disciplines.

Element f: Instruction and content are relevant to students and incorporate students' background and contextual knowledge.

Quality Standard II: Teachers establish a safe, inclusive and respectful learning environment for a diverse population of students.

Element a: Predictable classroom learning environment in which each student has a positive, nurturing relationship with caring adults and peers.

Element b: Commitment to and respect for diversity.

Element c: Engage students as individuals with unique interests and strengths.

Element d: Teaching adapted for the benefit of all students, including those with special needs, across a range of ability levels.

Element e: Work collaboratively with and provide feedback to students' families.

Element f: Learning environment characterized by acceptable student behavior, efficient use of time, and appropriate intervention strategies.

Quality Standard III: Teachers plan and deliver effective instruction and create an environment that facilitates learning for students.

Element a: Knowledge of current developmental science, the ways in which learning takes place, and the appropriate levels of intellectual, social, and emotional development of their students.

Element b: Instruction draws on results of student assessments, is aligned to academic standards, and advances students' content knowledge and skills.

Element c: Knowledge of current research on effective instructional practices to meet the developmental and academic needs of their students.

Element d: Integrate and utilize appropriate available technology to maximize student learning.

Element e: Communicate high expectations for all students and plan instruction that helps students develop critical-thinking and problem solving skills.

Element f: Students are provided opportunities to work in teams and develop leadership qualities.

Element g: Communicate effectively, making learning objectives clear and providing appropriate models of language.

Element h: Use appropriate methods to assess what each student has learned, including formal and informal assessments, and use results to plan further instruction.

Quality Standard IV: Teachers reflect on their practice.

Element a: Analyze student learning, development, and growth and apply what they learn to improve their practice.

Element b: Link professional growth to their professional goals.

Element c: Respond to a complex, dynamic environment.

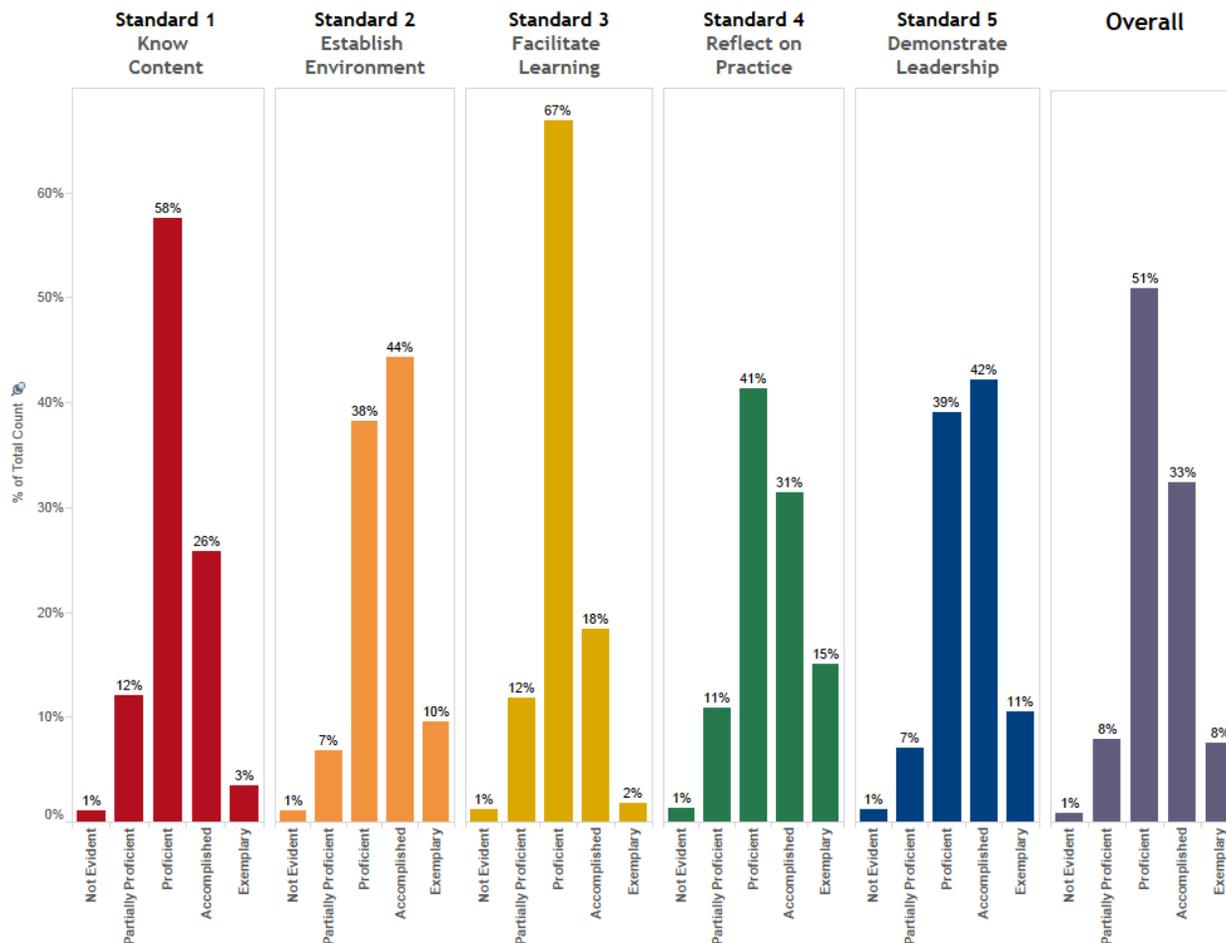
Figure 1 cont. Colorado Teacher Quality Standards and corresponding elements

Quality Standard V: Teachers demonstrate leadership.
Element a: Demonstrate leadership in their schools.
Element b: Contribute knowledge and skills to educational practices and the teaching profession.
Element c: Advocate for schools and students, partnering with students, families and communities as appropriate.
Element d: High ethical standards.

Section 1. Distributions of Standard and Overall Ratings

This section will review findings pertaining to the distributions of Standards 1 through 5, as well as the overall rating. Figure 2 presents the distributions of ratings across the five standards. In general, there are different distributions across the standards, indicating that the rubric measures different aspects of teacher practice and differentiates between teachers. The lack of uniformity also indicates that those who are evaluating teachers (e.g., principals, assistant principals, and teacher leaders) understand the differences between the standards and are able to apply the rubric to identify differences between teachers.

Figure 2. Standards and overall ratings distributions



Notes. Percentages may not add to 100 percent because of rounding.

Overall, pilot teachers received very high ratings on Standard 2, which pertains to establishing a safe and inclusive learning environment (this report will delve into each standard in the subsequent section). The finding that teachers received high ratings on Standard 2 can be interpreted in a number of ways. It is possible that teachers in the pilot districts truly have the most skill in establishing an environment that is safe and respectful and the ratings merely reflect this proficiency. It could also be that evaluators are not comfortable giving lower scores on this standard because it contains more sensitive text and references such as “nurturing relationship” and “respect for diversity.” Another explanation pertains to the professional practices rubric: it is possible that the rubric is more clear or explicit in Standard 2, and because evaluators understand each professional practice better they are more likely to give a teacher credit for displaying that practice. Conversely, the professional practices pertaining to Standard 2 may be more nebulous, and a lack of understanding could cause evaluators to give teachers the benefit of the doubt and give credit for displaying a behavior that they are unsure about.

In contrast, teachers received the lowest ratings on Standard 3, which is the standard that encompasses practices related to effective instruction and facilitating learning for all students. Interpretations for this finding are similar to those delineated above—this finding could be a true reflection of teacher skill, the result of evaluator comprehension of the rubric or comfort level assigning lower ratings, or a product of the text in the professional practice rubric (or some combination of the three). It is important to consider these multiple interpretations since it is the first year of the pilot and all findings are considered informative, but preliminary.

With regard to the overall ratings, 92 percent of teachers received a summative rating of proficient or higher with the largest number of educators receiving a rating of proficient. Thirty-three percent of teachers received a summative rating of accomplished and 8 percent earned the highest rating of exemplary. On the other end of the spectrum, 8 percent of teachers received a summative rating of partially proficient and 1 percent received the lowest rating of not evident.

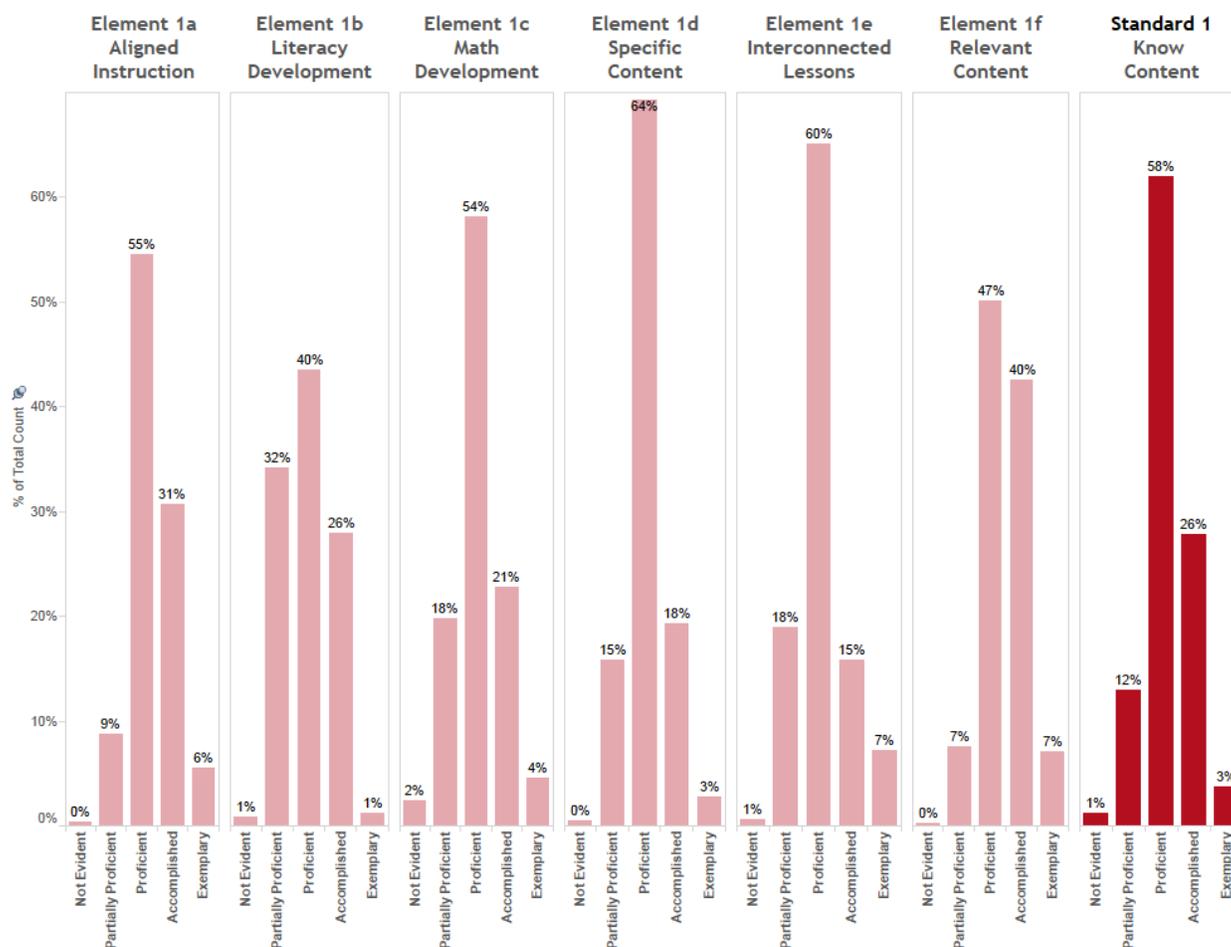
The standards are positively correlated with each other, indicating that although the rubric captures multiple aspects of teaching, these aspects are interrelated. The finding that they are positively correlated means that teachers who receive high ratings on one standard are more likely to receive high ratings on the other standards. Correlations between the standards are all moderately positive ($0.50 < \rho < 0.67$; calculated using Spearman’s rho, although Pearson correlation coefficients are nearly identical). As expected, each standard is strongly correlated to the overall rating ($0.73 < \rho < 0.80$). Reliability analyses also suggest that the ratings demonstrate high internal consistency, at a level consistent with typical ranges reported in large-scale standardized assessments (Cronbach’s $\alpha = 0.90$).

Section 2. Distributions of Element Ratings

This section explores the distributions of element ratings within each standard. The elements within Standard 1 (Know Content) are presented in Figure 3. Eighty-seven percent of teachers received a rating of proficient or higher on Standard 1 (with 29 percent receiving an accomplished or exemplary rating). Element 1b (Literacy Development) was one of the lowest rated elements in the professional practice rubric, with 67 percent of educators receiving a rating of proficient or higher.

The elements within Standard 1 are moderately correlated ($0.39 < \rho < 0.51$). Each element rating is strongly correlated to the overall standard rating ($0.63 < \rho < 0.71$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach’s $\alpha = 0.84$). All of the standards have reliability statistics that are at a level consistent with typical ranges reported in large-scale standardized assessments

Figure 3. Standard 1: Know Content - elements and summative rating

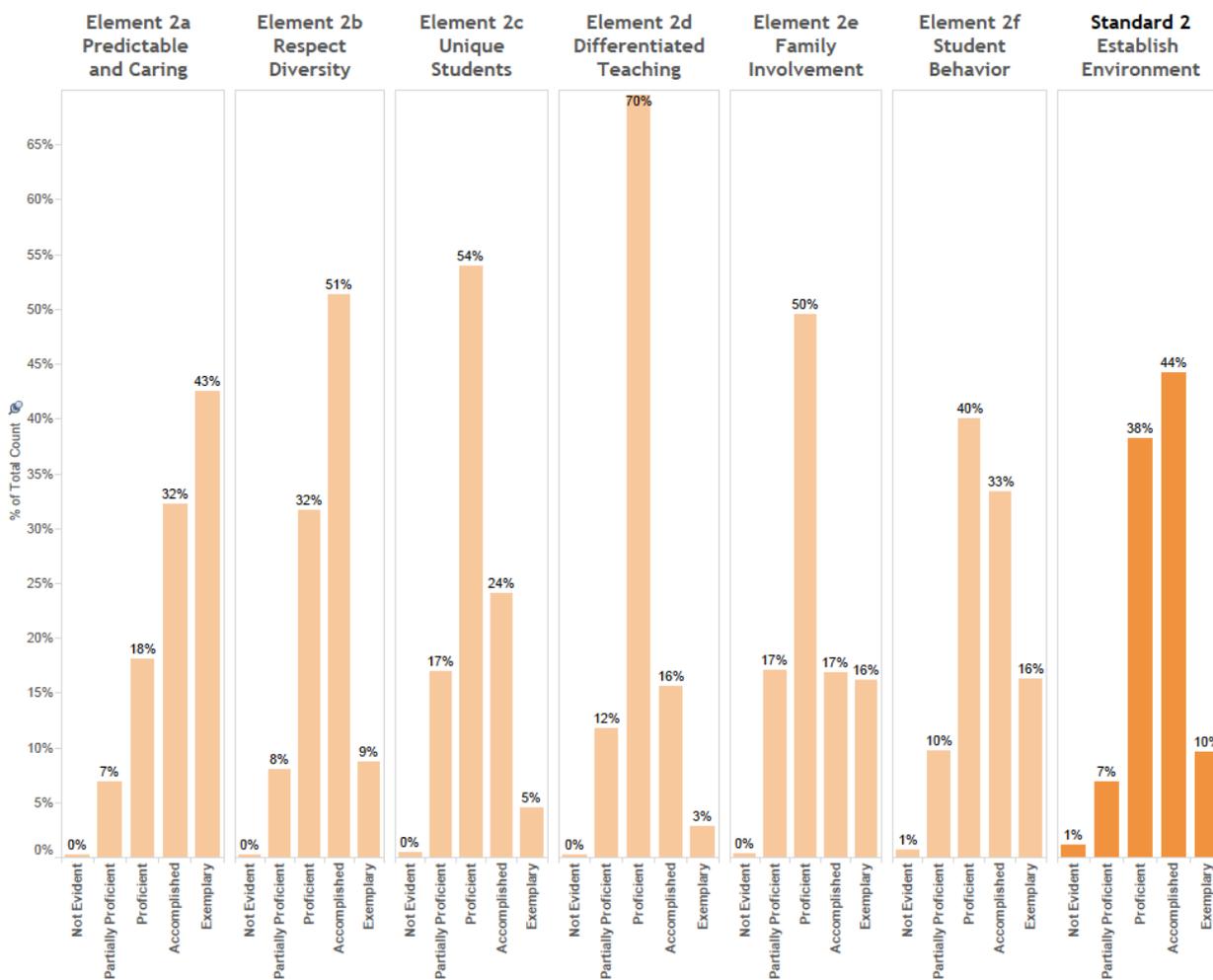


Notes. Percentages may not add to 100 percent because of rounding.

Represented in Figure 4, Standard 2 (Establish Environment) is one of the highest rated standards, with 92 percent of teachers receiving a rating of proficient or higher and 54 percent meeting the higher bar of accomplished or exemplary. Three of the highest rated elements on the rubric are in this standard: Elements 2a (Predictable and Caring), 2b (Respect Diversity), and 2f (Student Behavior).

The elements within Standard 2 are moderately correlated ($0.35 < \rho < 0.60$) and each element rating is strongly correlated to the overall standard rating ($0.61 < \rho < 0.73$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach's $\alpha = 0.83$).

Figure 4. Standard 2: Establish Environment - elements and summative rating

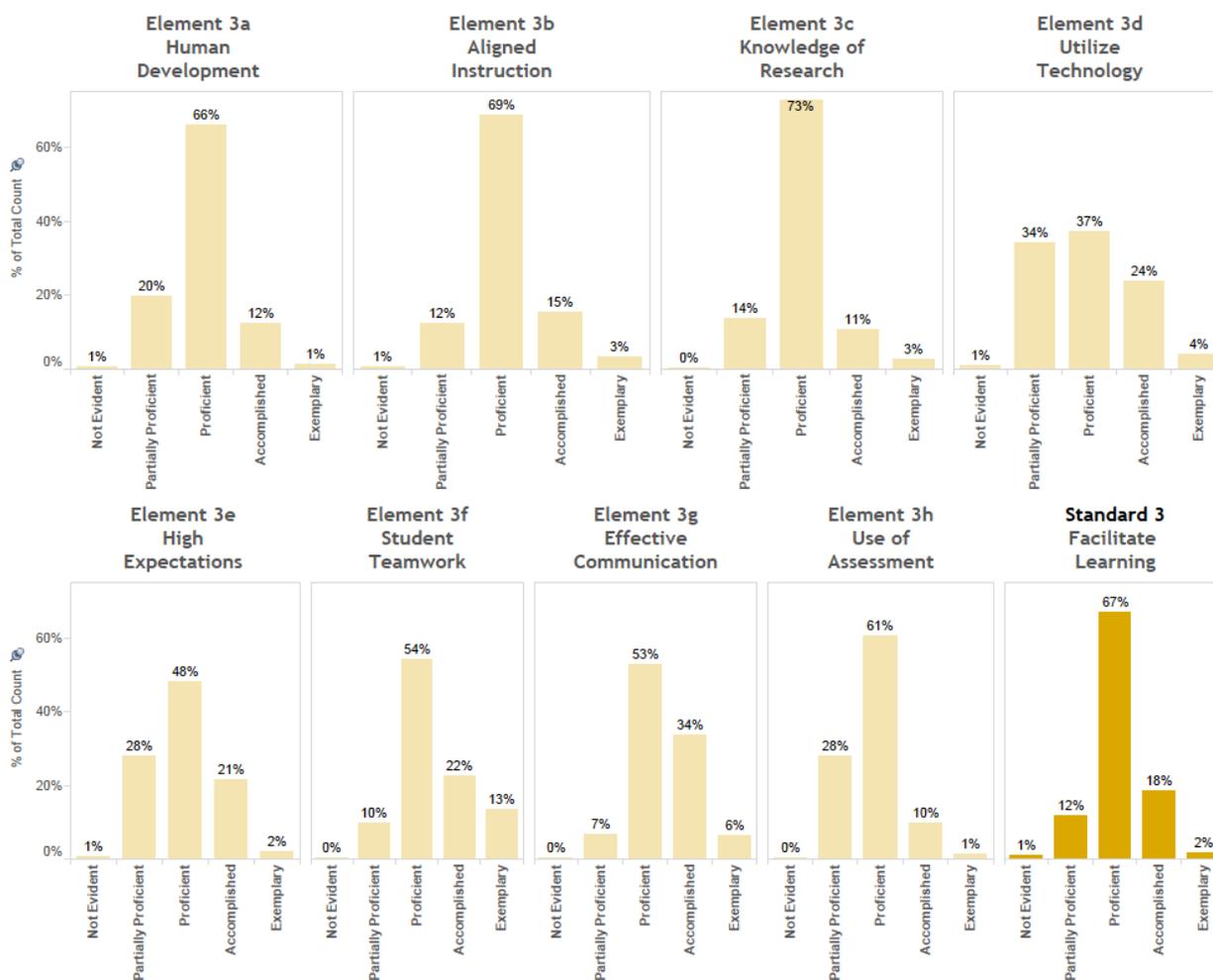


Notes. Percentages may not add to 100 percent because of rounding.

The elements and standard ratings for Standard 3 (Facilitate Learning) are shown in Figure 5. Standard 3 is the lowest rated standard with 87 percent of teachers receiving a rating of proficient or higher. Twenty percent of teachers were rated accomplished or exemplary. Four of the lowest rated elements on the rubric are in Standard 3: Elements 3a (Human Development), 3d (Utilize Technology), 3e (High Expectations), and 3h (Use of Assessment). Element 3h is the lowest rated element on the rubric, with only 11 percent of teachers reaching the high bar of accomplished or exemplary.

The elements within Standard 3 have low to moderate correlations ($0.28 < \rho < 0.52$). Each element rating is strongly correlated to the overall standard rating ($0.52 < \rho < 0.67$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach's $\alpha = 0.85$).

Figure 5. Standard 3: Facilitate Learning - elements and summative rating

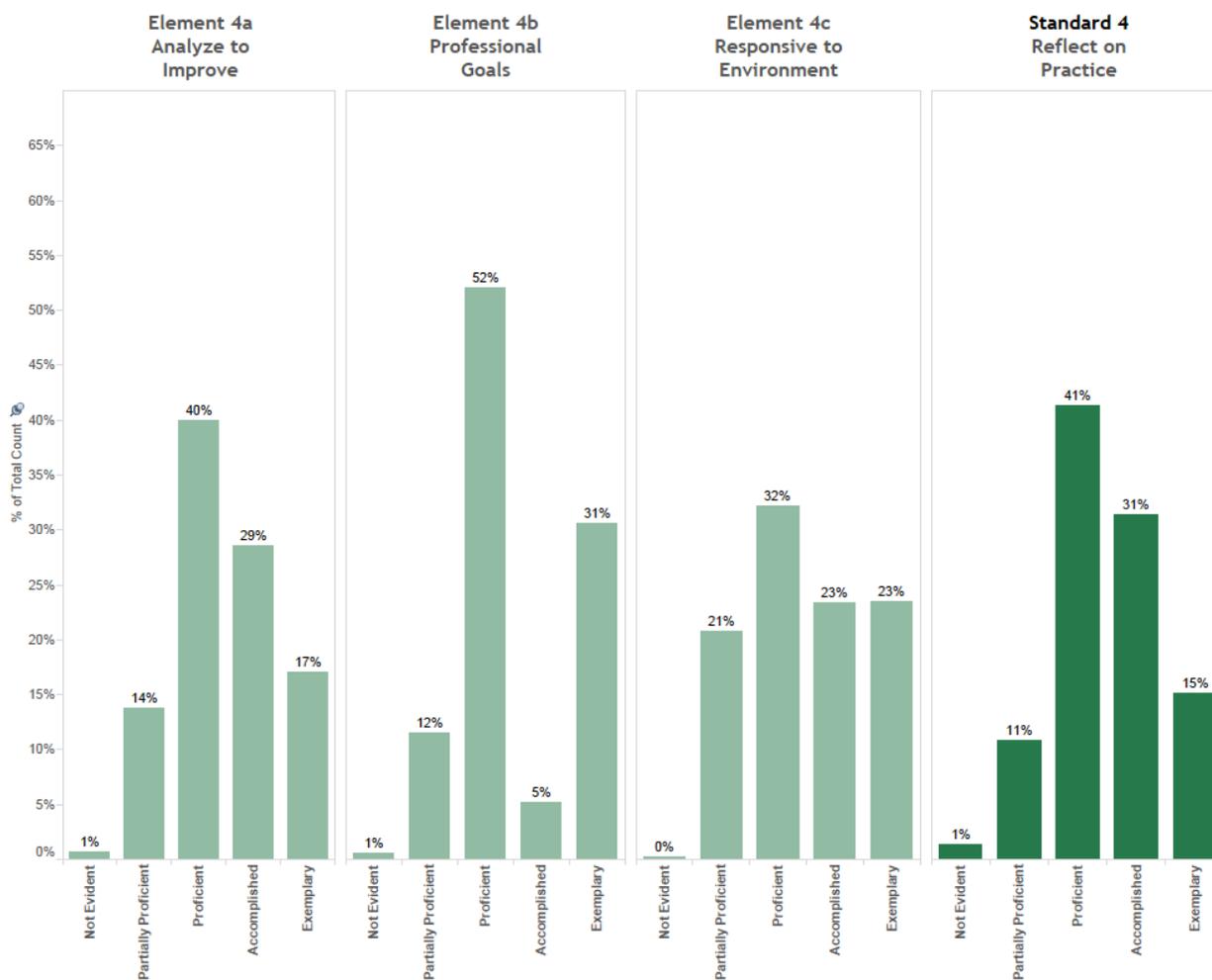


Notes. Percentages may not add to 100 percent because of rounding.

The distributions for the elements within Standard 4 (Reflect on Practice) are depicted in Figure 6. Eighty-seven percent of teachers received a rating of proficient or higher and 46 percent were rated accomplished or exemplary on the standard. Although Standard 4 is not the highest rated standard, more teachers received a summative rating of exemplary than on any other standard (15 percent received an exemplary rating).

The elements within Standard 4 are moderately correlated ($0.49 < \rho < 0.52$) and each element rating is strongly correlated to the overall standard rating ($0.75 < \rho < 0.80$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach's $\alpha = 0.75$).

Figure 6. Standard 4: Reflect on Practice - elements and summative rating

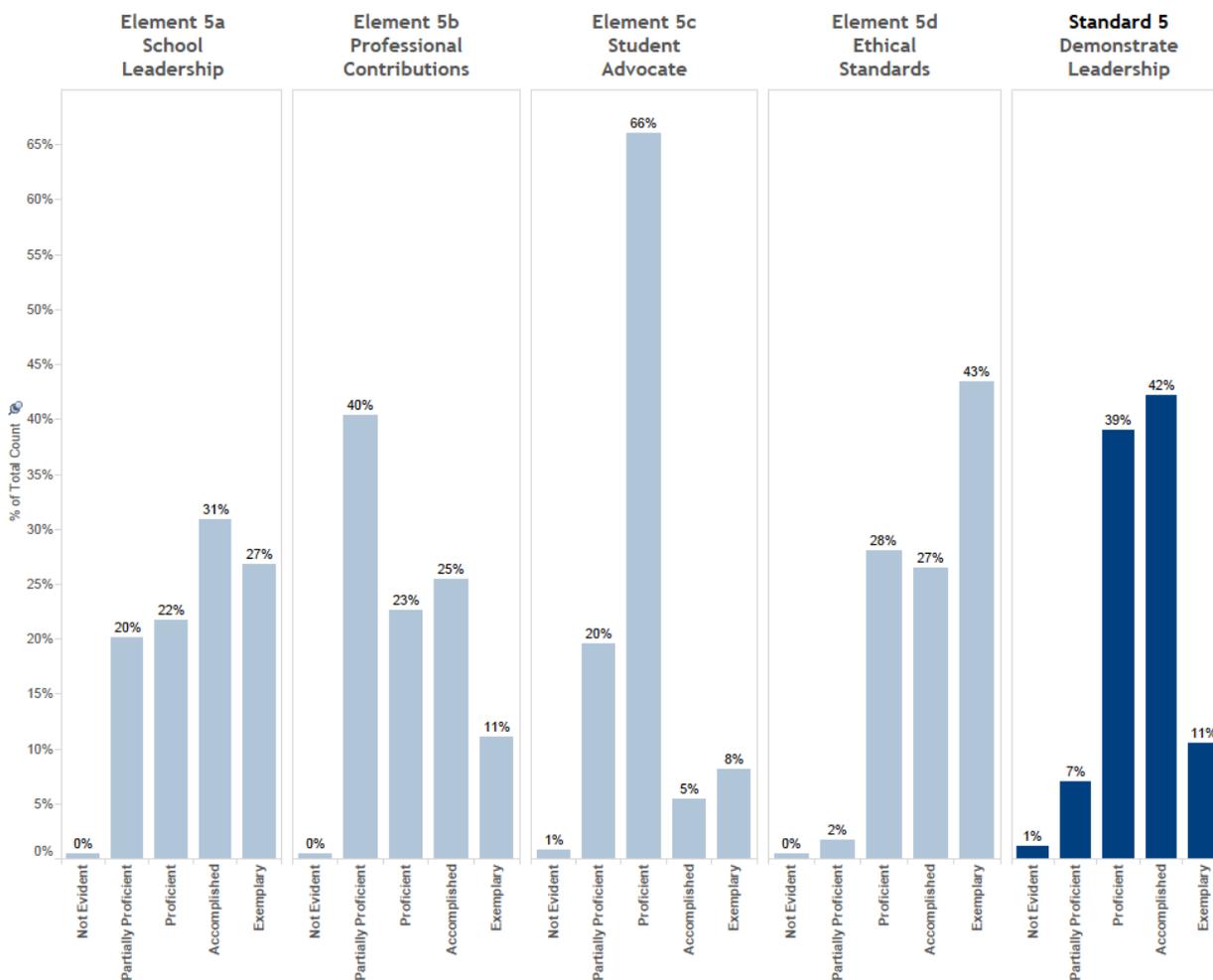


Notes. Percentages may not add to 100 percent because of rounding.

The other highest rated standard on the rubric is Standard 5 (Demonstrate Leadership) with 92 percent of teachers receiving a rating of proficient or higher and 53 percent meeting the higher bar of accomplished or exemplary (illustrated in Figure 7). The highest rated element on the rubric is Element 5d (Ethical Standards), with a notable 70 percent of teachers rated accomplished or exemplary. Element 5a (School Leadership) is also one of the highest rated elements.

The elements within Standard 5 have low to moderate correlations ($0.28 < \rho < 0.58$). Each element rating is strongly correlated to the overall standard rating ($0.61 < \rho < 0.79$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach’s $\alpha = 0.75$).

Figure 7. Standard 5: Demonstrate Leadership - elements and summative rating



Notes. Percentages may not add to 100 percent because of rounding.

Section 3. Ratings Throughout the School Year

The professional growth and evaluation cycle in the teacher state model evaluation system starts at the beginning of the school year when teachers conduct a self-assessment of their practice. The self-assessment helps educators identify areas of strength and areas for professional growth. The areas for growth can then be worked on throughout the school year. Similarly, teachers should receive mid-year ratings and have a mid-year conversation with their evaluator in the middle of the school year. This is another opportunity to identify areas of strength and areas that need improvement. Ideally, by the time the final evaluation occurs at the end of the year, teachers have had the opportunity to focus on a couple areas of growth and increase their ratings in those areas. At a minimum, it is expected that teachers would maintain rating levels by the end of the year. To examine growth in ratings throughout the school year, CDE looked at the percentage of teachers who maintained or improved their overall rating by the end-of-year evaluation as well as the correlations between ratings at different points in the school year. As shown in Figure 8, the large majority of teachers either maintained or improved their ratings by the end-of-year (or final) evaluation.

Figure 8. Relationship between self-assessment, mid-year, and final evaluation ratings

	Percent of teachers who maintained performance category on final evaluation	Percent of teachers who moved up a performance category on final evaluation ^a	Percent of teachers who maintained or moved up a performance category on final evaluation	Correlation with final evaluation rating ^b
Self-assessment rating	55%	34%	89%	.52
Mid-year rating ^c	72%	28%	100%	.70

^a Includes teachers who maintained their exemplary rating, since that is the highest performance category.

^b Correlation coefficients reported are Spearman rho.

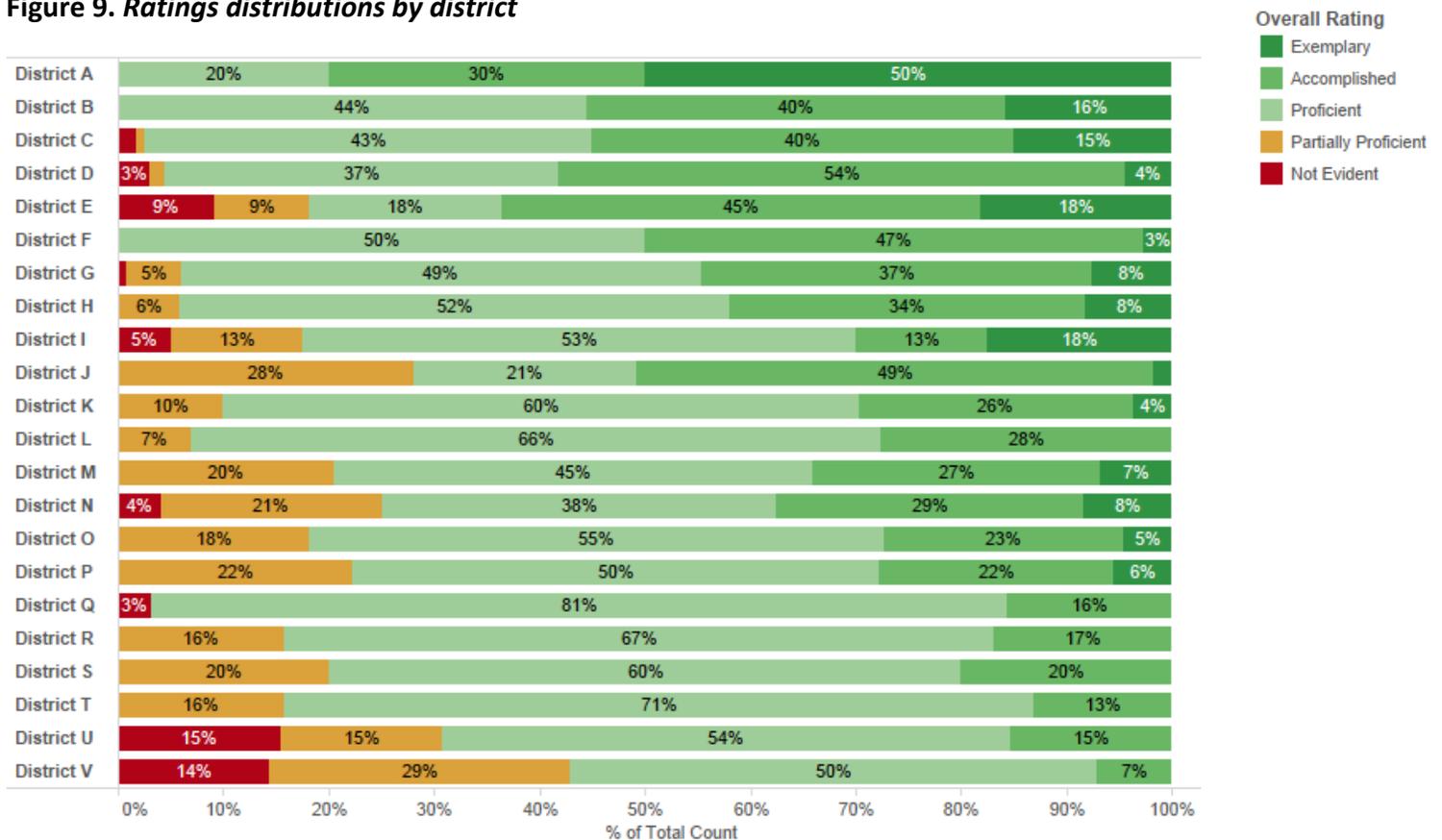
^c The mid-year rating analyses involve a smaller sample of teachers because districts did not always submit mid-year ratings (n=735).

Section 4. Ratings Distributions by District

As stated previously, 26 districts piloted the teacher model evaluation system and 25 of those districts submitted final evaluation ratings for the districts' teachers. Differences between the overall ratings distributions in these districts are illustrated in Figure 9. District names and sample sizes have been removed to protect district confidentiality, and districts with fewer than 10 teachers participating in the pilot have been removed.

There are large differences between districts in the percent of teachers in each performance category. Comparing at the extremes, 100 percent of teachers in District A are proficient or higher while 57 percent of teachers in District V are proficient or higher. It is important to remember that these findings are preliminary and that multiple interpretations are possible. For example, it is possible that the teachers in District A are truly the most proficient teachers in the pilot and therefore received the highest ratings. However it is also possible that these ratings are a function of the evaluators and that evaluators in District A are not comfortable assigning ratings lower than proficient. It is also possible that evaluators in District A are less familiar with the professional practice rubric and are assigning so many exemplary ratings because of their lack of understanding of what exemplary practices should look like.

Figure 9. Ratings distributions by district



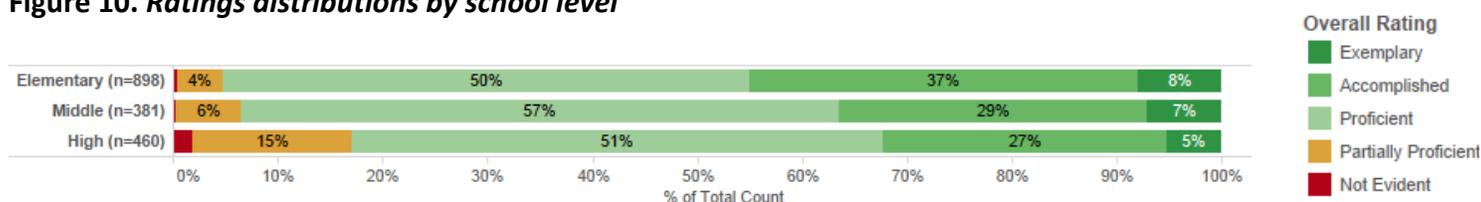
Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by district, meaning that the distribution of overall ratings varies as a function of the district.

Section 5. Ratings Distributions by Teacher Employment and Demographic Characteristics

This section examines differences in teacher performance category by employment and demographic characteristics. Note that the teacher employment and demographic data are from the official Human Resources (HR) collection that districts submit to CDE every year. Because of the lag in reporting, the measures used in this section are from the 2011-2012 CDE HR collection.¹ Figures 10-18 present ratings based on a range of characteristics.

Starting with teachers' employment characteristics, Figure 10 displays the differences between elementary, middle, and high school teachers. There are statistically significant differences in teachers' overall ratings based on the school level. Elementary teachers received the highest ratings, followed by middle school teachers and high school teachers (all differences between groups are statistically significant).

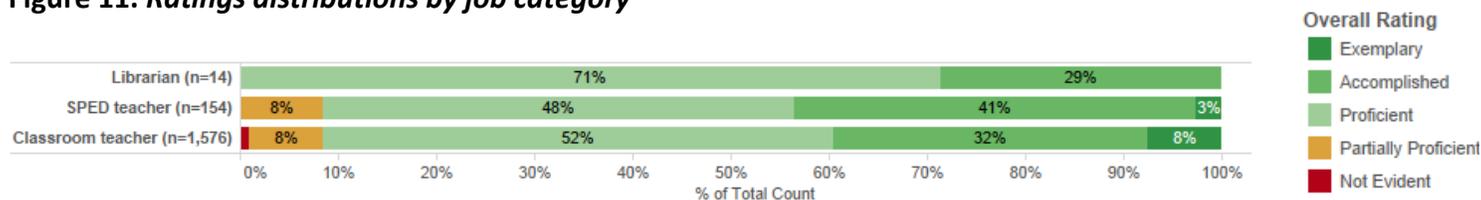
Figure 10. Ratings distributions by school level



Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by school level, meaning that the distribution of overall ratings varies as a function of the school level.

In contrast to school level, there are no statistically significant group differences by job category (presented in Figure 11). Although librarians did receive higher ratings, they are a small sample and therefore it is inappropriate to make conclusions about the perceived differences between librarians and classroom teachers. Likewise, special education teachers did receive higher ratings than classroom teachers, but the differences are not large enough to be statistically significant.

Figure 11. Ratings distributions by job category

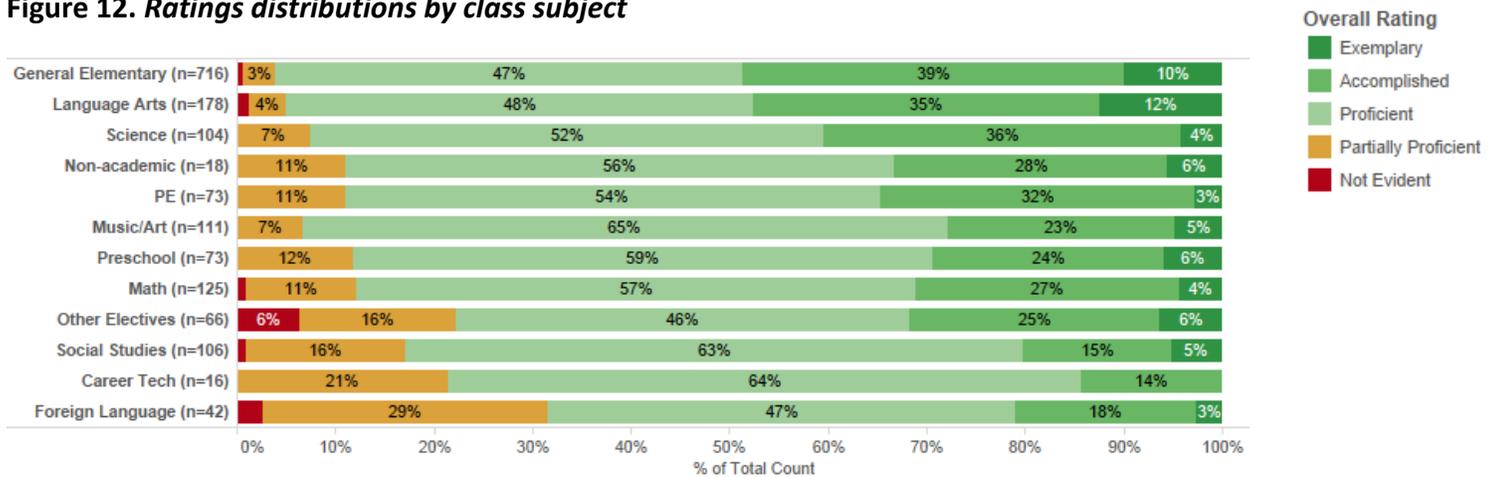


Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are no statistically significant group differences by job category.

¹ Because the HR measures are from 2011-2012, approximately 150 teachers who were new to the school or district in 2012-2013 will not have data and are not represented in these analyses. Also note that the findings reported exclude groups with fewer than 10 educators.

Limiting the analyses to classroom teachers, there are statistically significant differences between teachers' ratings based on the subject they teach (depicted in Figure 12). The differences indicate that general elementary teachers receive the highest ratings, which is consistent with findings related to the school level.

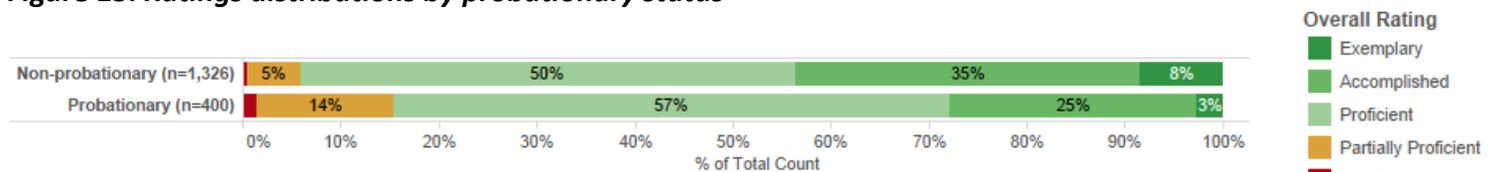
Figure 12. Ratings distributions by class subject



Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by class subject, meaning that the distribution of overall ratings varies as a function of the class subject.

The differences in teachers' overall ratings based on probationary status are statistically significant. Teachers with non-probationary status receive higher ratings, as depicted in Figure 13.

Figure 13. Ratings distributions by probationary status



Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by probationary status, meaning that the distribution of overall ratings varies as a function of teachers' probationary status.

There are also statistically significant relationships between a teacher’s years of experience and overall rating (see Figure 14; rho = 0.22) and a teacher’s salary and overall rating (see Figure 15; rho = 0.21). On average, teachers with higher ratings have more years of experience and earn a higher salary.

Figure 14. Average years of experience for each performance category

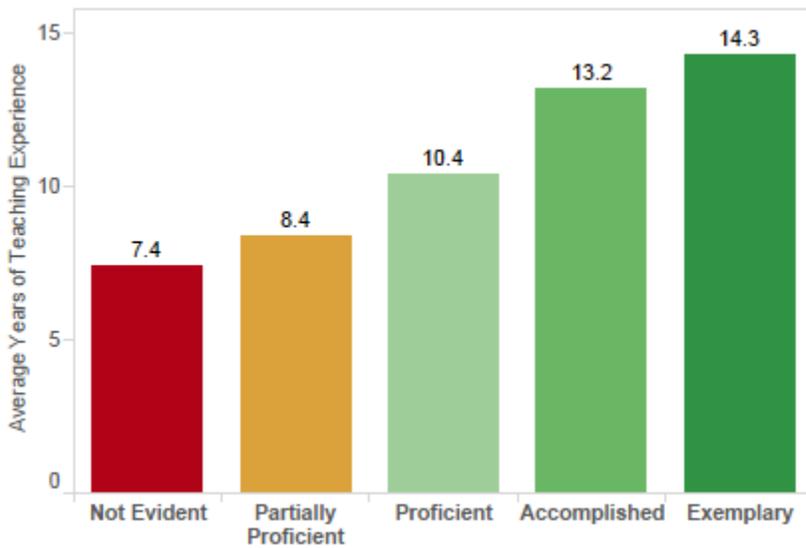
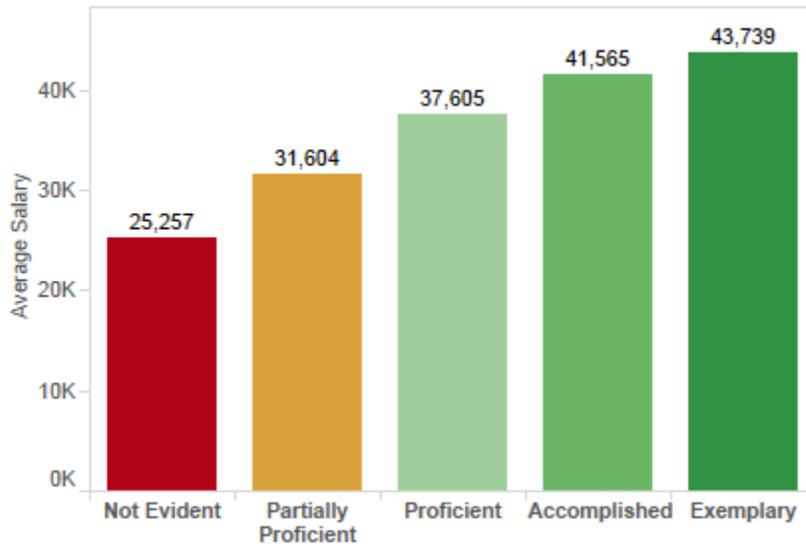
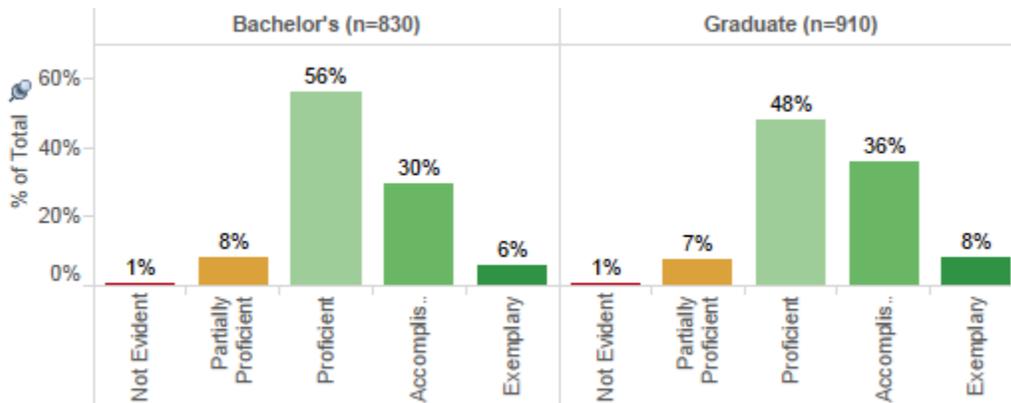


Figure 15. Average salary for each performance category



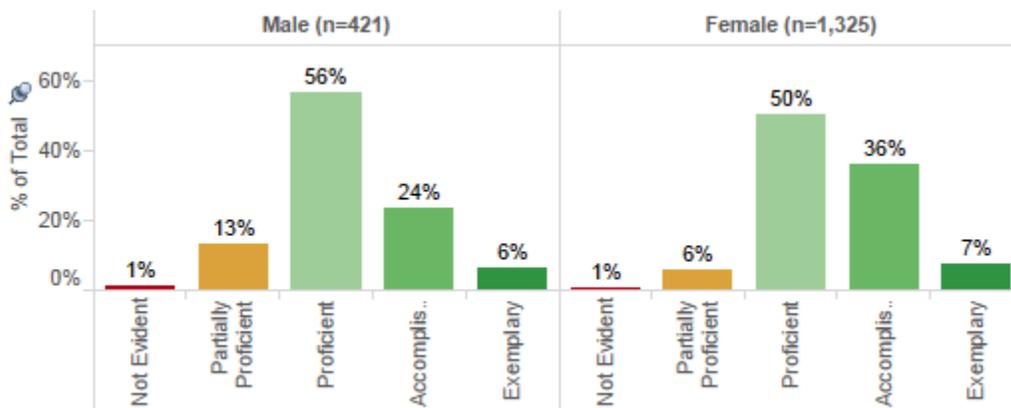
The final group differences examined pertain to teachers' education level, gender, and ethnicity. Teachers with a graduate degree receive higher ratings than teachers with a bachelor's degree (see Figure 16). Females receive higher ratings than males (see Figure 17) and white teachers receive higher ratings than Latino teachers (see Figure 18); all other racial/ethnic groups had fewer than 10 teachers in the group and are therefore not reported.

Figure 16. Ratings distributions based on teachers' education level



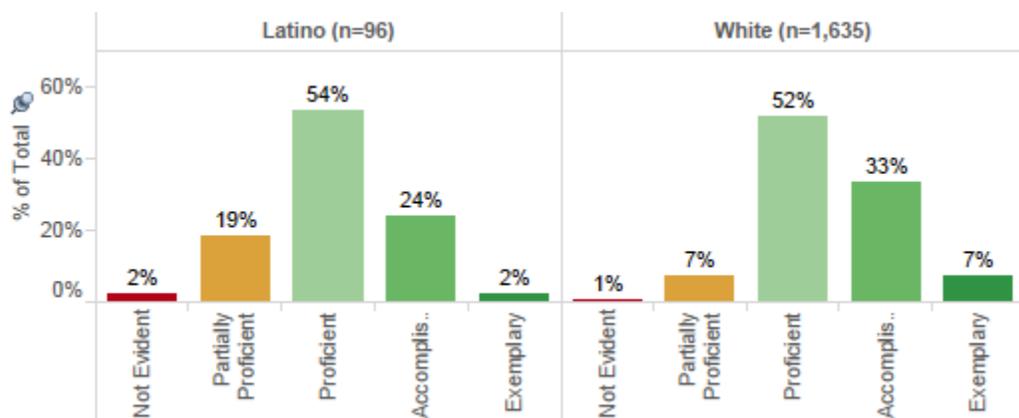
Notes. There are statistically significant group differences by teachers' education level, meaning that the distribution of overall ratings varies as a function of education level.

Figure 17. Ratings distributions based on teachers' gender



Notes. There are statistically significant group differences by teachers' gender, meaning that the distribution of overall ratings varies as a function of gender.

Figure 18. Ratings distributions based on teachers' ethnicity



Notes. There are statistically significant group differences by teachers' ethnicity, meaning that the distribution of overall ratings varies as a function of ethnicity.

Summary and Next Steps

Overall, preliminary analyses from the 2012-2013 pilot of the Colorado State Model Evaluation System for teachers indicate that the professional practice rubric captures multiple aspects of teaching and differences in teacher practice. In sum, these results lend credibility to the Colorado Teacher Quality Standards and the elements of which they are comprised. Educators receive the highest ratings on elements related to establishing a caring and respectful classroom environment and demonstrating leadership. They receive the lowest ratings on elements related to effective instruction and facilitating learning for all students.

One of the most promising findings relates to the progression of teachers' practice during the school year. The model evaluation system is built on the belief that teachers who receive clear and frequent feedback about their teaching will improve their practice and therefore further impact student learning. Preliminary findings indicate that the majority of teachers maintained or improved their practice through the course of the school year.

This report identifies multiple factors that are associated with a teacher's final evaluation rating, including district, school, and teacher-level factors (e.g., school level and teacher gender). Additional quantitative and qualitative analyses will explore what may contribute to such differences, whether these group differences are a function of human judgment or a function of the rubric, and the implications of these differences on students in classrooms.

CDE will continue to run reliability and validity analyses on the 2012-2013 teacher model evaluation system data, including examining relations with other measures of classroom success. Substantive changes were made to the teacher rubric for the 2013-2014 school year based on the findings from this analysis and qualitative data (from surveys and focus groups). Rubric modifications include changing the not evident category to basic (as well as adjusting the scoring for that performance category), removing redundant professional practices, removing most non-observable professional practices from Standards 1, 2, and 3, and shortening the rubric by six pages. All of these changes will impact future findings so CDE will continue to collect and analyze quantitative and qualitative data in subsequent school years.