COLORADO
Department of Education

Colorado State Model Evaluation System for Teachers

2013-2014 Pilot Report

Introduction

Senate Bill 10-191, passed in 2010, restructured the way teachers are supported and evaluated in Colorado. The ultimate goal is ensuring college and career readiness for all students, which is greatly impacted by the effectiveness of the teachers and leaders in schools. To support this effort, the Colorado Department of Education (CDE) developed a model system as an option for districts to use in implementing the new evaluation requirements for educators. Currently, 160 districts have opted to utilize the Colorado State Model Evaluation System for Teachers.

The Colorado State Model Evaluation System for Teachers was first piloted in 26 school districts of varying size and location during the 2012-2013 school year (25 of those districts piloted again in 2013-2014). The 2012-2013 pilot report presented findings from the 1,900 teachers in 25 districts that submitted final evaluation ratings (note that there is also a 2012-2013 report of the Colorado State Model Evaluation System for Principals). The current report presents findings from the 2013-2014 pilot of the teacher model evaluation system, including professional practice data from 3,436 teachers in 23 (out of 25 participating) districts (a separate report on the principal model evaluation system is forthcoming). When appropriate and informative, CDE will report comparisons between years one and two of the teacher model evaluation system pilot.

As with the 2012-2013 findings, the 2013-2014 findings should be considered preliminary for the following reasons:

1. Although 2013-2014 was the second year teachers and their evaluators experienced the teacher model evaluation system, each school and district is participating in a continuous improvement cycle with regard to the new evaluation process. Some educators received more training on the evaluation system than educators in other districts or other schools. Evaluators implement the evaluation process with varying levels of fidelity given their own training and time constraints. Training and implementation fidelity can affect ratings in different ways. It is important to give schools and districts time to improve their processes. CDE will be able to make more assertions about findings once there is additional evidence for reliability and implementation fidelity. To support district efforts and growth, CDE continues to provide resources to assist districts, including an online system to help with inter-rater agreement (Elevate Colorado) and an online platform to organize professional growth and evaluation information (Colorado State Model Performance Management System).

2. The Colorado State Model Evaluation System for Teachers has its own continuous improvement cycle and CDE is conducting ongoing analyses of quantitative and qualitative data to make changes. 2012-2013 was the first year CDE was able to collect pilot data and conduct such analyses and changes were made prior to the 2013-2014 school year (e.g., shortening the rubric as well as substantive changes to the lowest performance category). Findings from the 2013-2014 pilot have informed changes for the 2014-2015 school year.

3. Educators reported during both years of the pilot that the conversations that result from ratings on the professional practice rubric are productive professional growth conversations. At the same time, these conversations can be challenging, and evaluators understandably find it difficult to identify and discuss areas that need improvement. Lack of comfort with delivering difficult feedback may lead to more positive ratings being assigned.

Summary of Key Findings

- The distributions of teacher ratings, at the element and Quality Standard levels, indicate that the professional practice rubric captures multiple aspects of teaching as well as differences in teacher practice. In general, there is less variability in year two of the pilot than was seen in year one (i.e., there is more clustering in the middle performance categories in year two).
- The variability in the ratings distributions also suggests that teacher evaluators are able to differentiate between teachers and assign ratings in a meaningful way.
- Teachers receive the highest ratings on Standard 4 (Reflect on Practice) and the lowest ratings on Standard 3 (Facilitate Learning).
- The majority of teachers maintained or improved their ratings in year two of the pilot.
- Teacher ratings vary based on the district, subject taught, probationary status, experience, and teacher demographics.
- Teacher ratings also vary based on school characteristics including the school level, School Performance Framework (SPF) rating, and student demographics.

Colorado Teacher Quality Standards

Before reviewing findings from the second year of the pilot of the teacher model evaluation system, it is important to consider what exactly comprises the Colorado Teacher Quality Standards, which are the foundation of the professional practice rubric. Note that the rubric measures Standards 1 through 5 (summarized in Figure 1). Standard 6, which pertains to teacher responsibility for student academic growth, was piloted during the 2013-2014 school year. Findings related to Standard 6 will be reported separately.

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.

Figure 1 cont. *Colorado Teacher Quality Standards and corresponding elements*

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.

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.

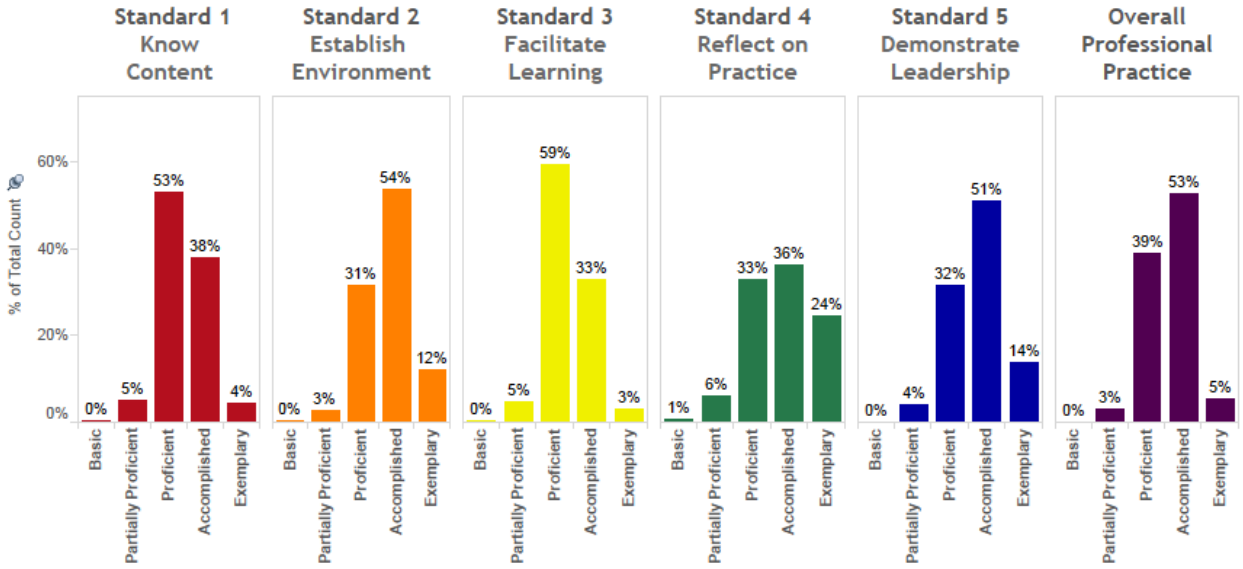
Contents of Report

The subsequent sections of this report focus on the results of the teacher model evaluation system year two pilot analyses. The sections are as follows:

Section 1. Distributions of Standard and Overall Ratings

The results section of this report begins with a description of the distributions of Quality Standards 1 through 5, as well as the overall professional practice rating, presented in Figure 2 (the report will delve into each standard in the subsequent section). Similar to year one of the pilot, there are different distributions across the standards, indicating that the rubric accomplishes the important tasks of measuring different aspects of teacher practice and differentiating between teachers. The lack of uniformity also indicates that those who are evaluating teachers (e.g., principals and assistant principals) understand the differences between the standards and are able to apply the rubric to identify differences between teachers. In comparison to year one, in year two there is more clustering (i.e., larger numbers of teachers) in the proficient and accomplished performance categories.

Figure 2. ***Standard and overall ratings distributions***



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

Pilot teachers received the highest ratings on Standard 4, which captures teachers' reflection on their practice and focus on professional growth[1]. In year one of the pilot, teachers received the highest ratings on Standard 2, which pertains to establishing a safe and inclusive learning environment. In contrast, teachers received the lowest ratings on Standard 3 in years one and two. Standard 3 is the standard that encompasses practices related to effective instruction and facilitating learning for all students.

With regard to overall professional practice ratings, 97 percent of teachers received a summative rating of proficient or higher with the largest number of educators receiving a rating of accomplished (compared to the distribution in year one where 92 percent of teachers received a summative rating of proficient or higher and proficient was the most common rating). Thirty-nine percent of teachers received a summative rating of proficient, 53 percent received a summative rating of accomplished, and 5 percent earned the highest rating of exemplary. On the other end of the

[1] Note that the distinction of "highest" and "lowest" rated standards and elements is based on an average across all rating categories, which takes into account the number of teachers in each category.

spectrum, 3 percent of teachers received a summative rating of partially proficient and .1 percent of teachers received the lowest rating of basic (which is why this group shows up as 0 percent in the graph).

All five standards are positively correlated with each other, indicating that the rubric captures multiple though interrelated aspects of teaching. 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. The standards are all moderately to strongly correlated[2] with each other ($0.49 < \rho < 0.69$; calculated using Spearman's rho, although Pearson correlation coefficients are nearly identical) and each standard is strongly correlated to the overall professional practice rating ($0.68 < \rho < 0.74$). 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 $\alpha = 0.87$).

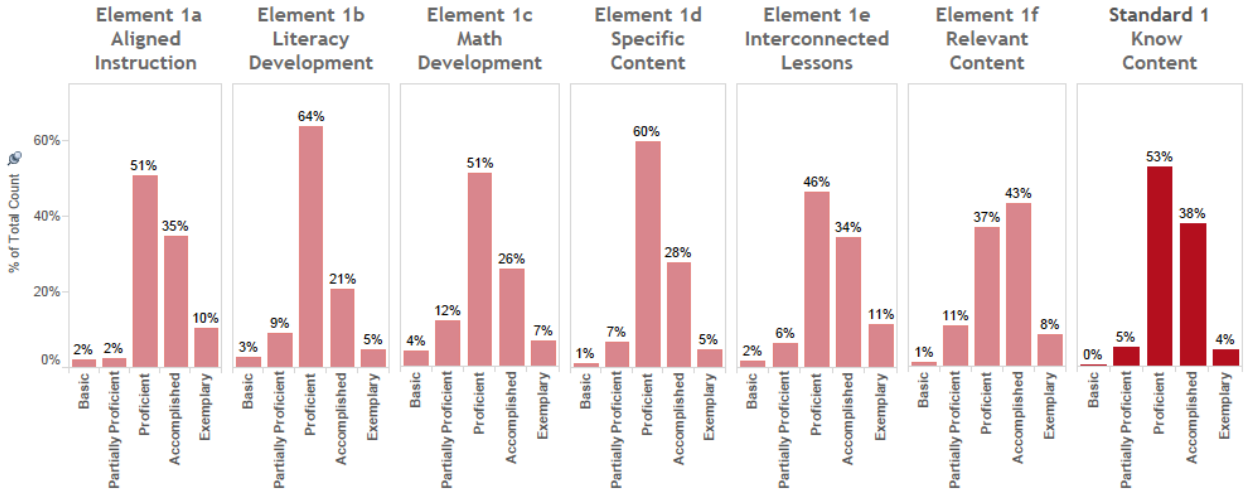
[2] Correlation coefficients indicate the strength of the relationship between two measures; a value of 0 indicates no relationship and a value of 1 indicates a perfect positive relationship (while a value of -1 indicates a perfect negative relationship). General guidelines for interpreting the value of the coefficient are: a correlation coefficient under .3 indicates a weak relationship, .3-.49 indicates a moderate relationship, and .5 and above indicates a strong relationship.

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. Ninety-five percent of teachers received a rating of proficient or higher on Standard 1 (with 42 percent receiving an accomplished or exemplary rating). Element 1b (Literacy Development) is one of the lowest rated elements on the professional practice rubric, with 88.5 percent of educators receiving a rating of proficient or higher (conversely, 11.5 percent of educators received a rating below proficient). Element 1b was also one of the lowest rated elements in pilot year one, when only 67 percent of educators receiving a rating of proficient or higher. Although the average across all rating categories may mask some of the lower ratings, higher percentages of teachers received below proficient ratings on Element 1c (Math Development; 16 percent below proficient) and Element 1f (Relevant Content; 12 percent below proficient).

The elements within Standard 1 are moderately to strongly correlated ($0.36 < \rho < 0.53$). Each element rating is strongly correlated to the overall standard rating ($0.59 < \rho < 0.70$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach's $\alpha = 0.83$). 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***

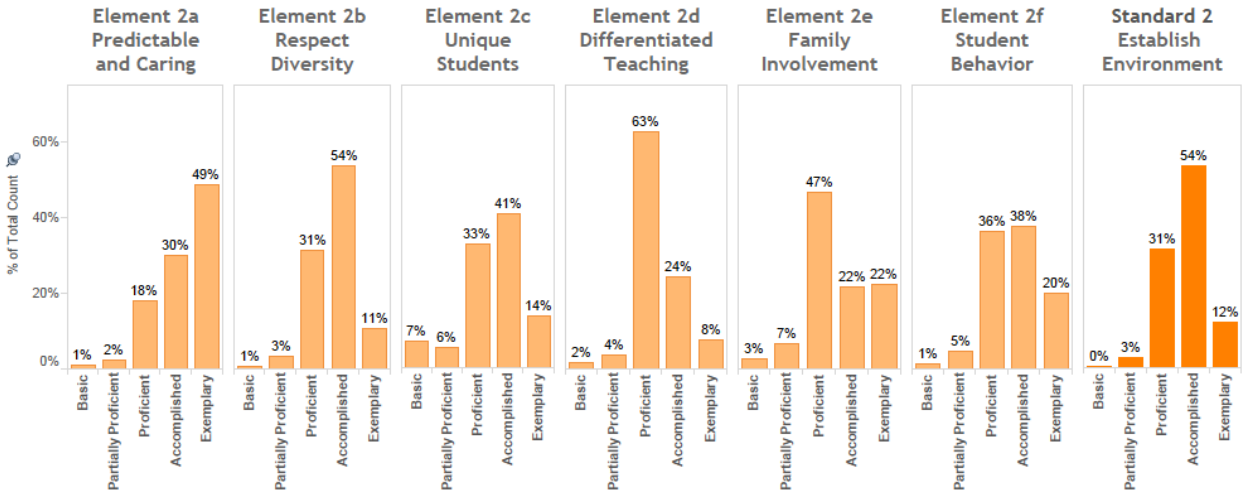


Notes. Percentages may not add to 100 percent because of rounding.

Standard 2 (Establish Environment) is depicted in Figure 4, with 97 percent of teachers receiving a rating of proficient or higher and 66 percent meeting the higher bar of accomplished or exemplary. Element 2a (Predictable and Caring) is one of the highest rated elements on the rubric, as it was in year one. Conversely, 13 percent of teachers received a rating below proficient on Element 2c (Unique Students).

The elements within Standard 2 are moderately to strongly correlated ($0.38 < \rho < 0.60$) and each element rating is strongly correlated to the overall standard rating ($0.61 < \rho < 0.71$). 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***

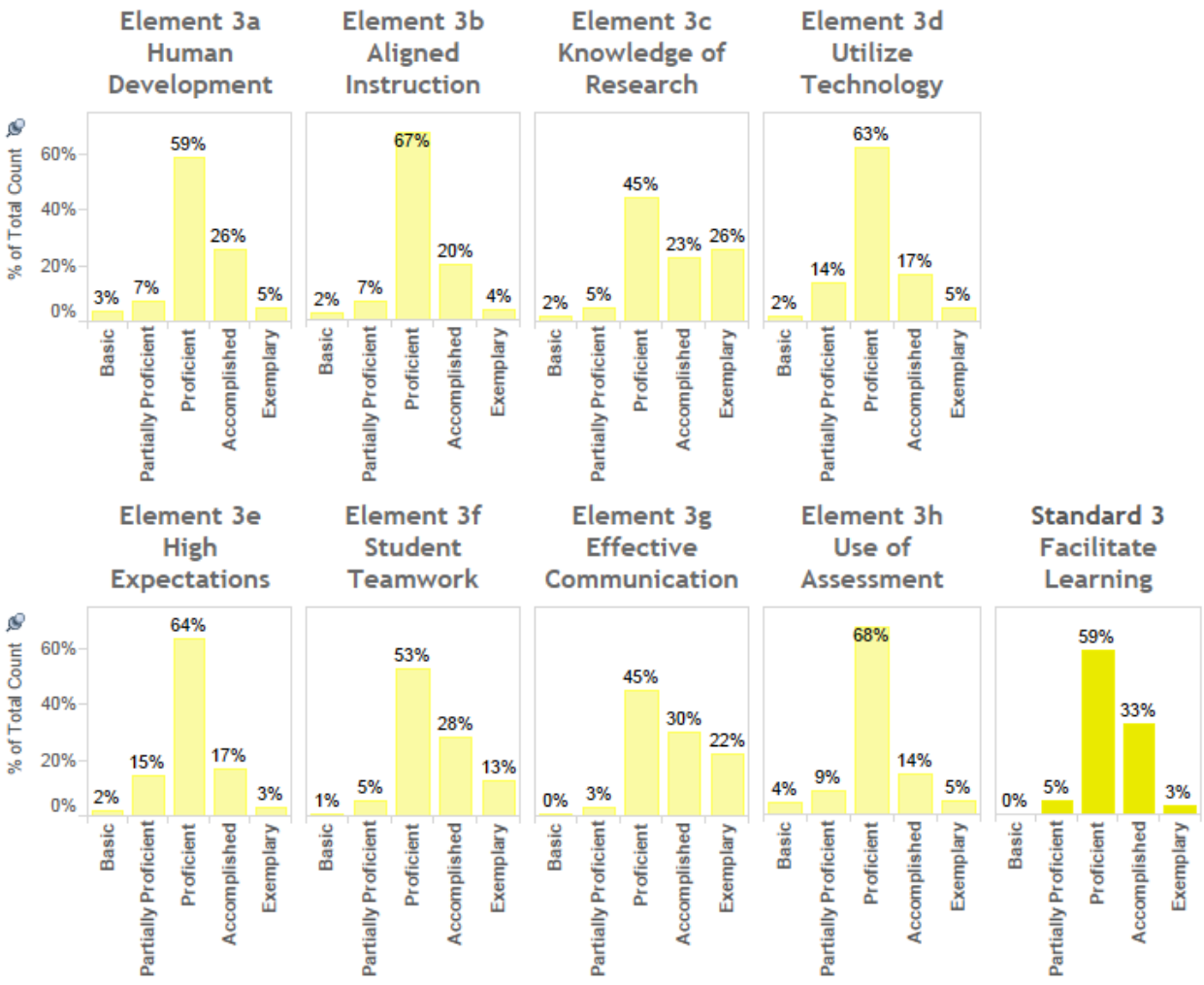


Notes. Percentages may not add to 100 percent because of rounding.

The distributions for the elements within Standard 3 (Facilitate Learning) are shown in Figure 5. Similar to year one, Standard 3 is the lowest rated standard with 95 percent of teachers receiving a rating of proficient or higher (in year one, 87 percent of teachers received a rating of proficient or higher). Thirty-six percent of teachers were rated accomplished or exemplary. Three of the lowest rated elements on the rubric are in Standard 3: 3d (Utilize Technology; 16 percent rated below proficient), 3e (High Expectations; 17 percent rated below proficient), and 3h (Use of Assessment; 13 percent rated below proficient). All three of these elements were among the lowest rated in year one as well.

The elements within Standard 3 have correlations ranging from weak to strong ($0.27 < \rho < 0.53$). Each element rating is moderately to strongly correlated to the overall standard rating ($0.47 < \rho < 0.70$). 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***

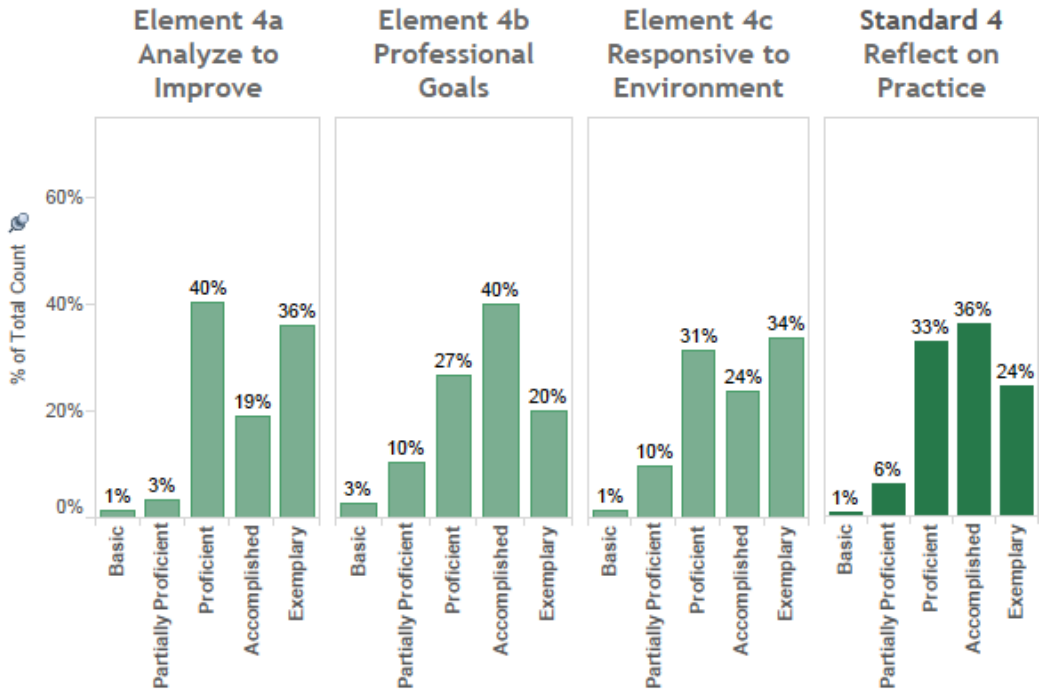


Notes. Percentages may not add to 100 percent because of rounding.

Standard 4 (Reflect on Practice) is the highest rated standard (represented in Figure 6). Ninety-three percent of teachers received a rating of proficient or higher and 60 percent were rated accomplished or exemplary on the standard (with a notable 24 percent earning the highest rating of exemplary). Elements 4a (Analyze to Improve) and 4c (Responsive to Environment) are among the highest rated elements on the rubric. Although Element 4c is one of the highest rated elements, 11 percent of teachers did receive a below proficient rating on this element. Likewise, 13 percent of teachers received a rating below proficient on Element 4b (Professional Goals).

The elements within Standard 4 are strongly correlated ($0.52 < \rho < 0.54$) and each element rating is strongly correlated to the overall standard rating ($0.77 < \rho < 0.82$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach's $\alpha = 0.77$).

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

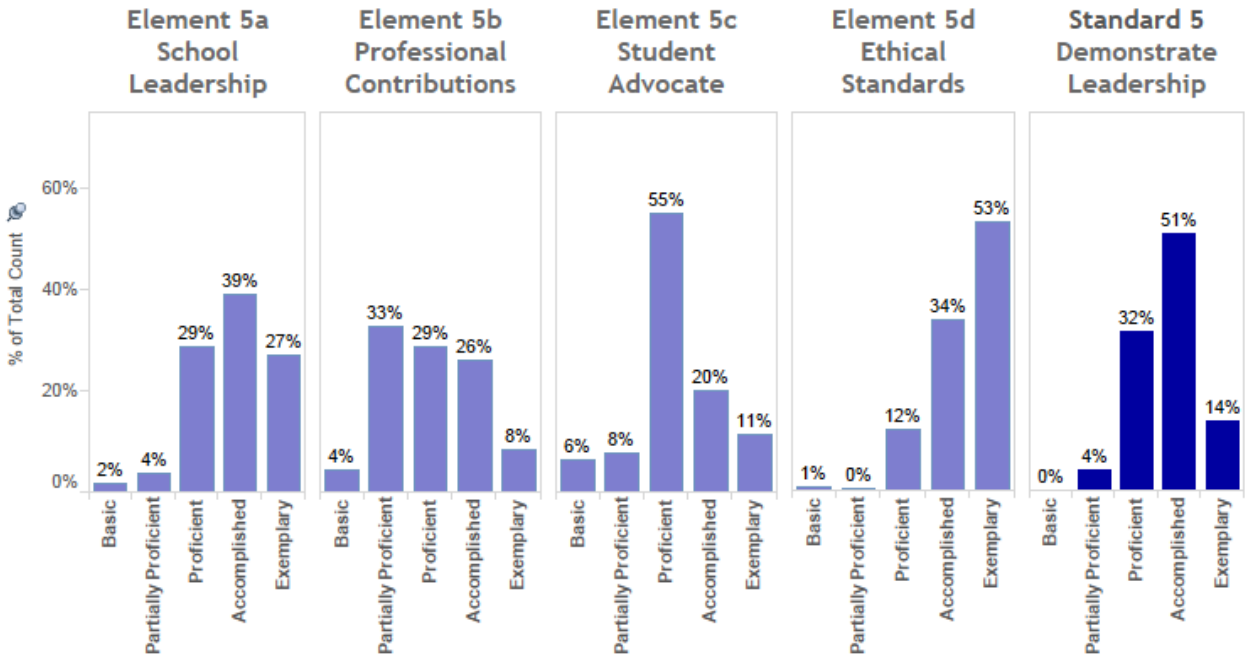


Notes. Percentages may not add to 100 percent because of rounding.

The distributions for the elements within Standard 5 (Demonstrate Leadership) are depicted in Figure 7. Ninety-seven percent of teachers received a rating of proficient or higher and 65 percent met the higher bar of accomplished or exemplary. The highest rated element on the rubric is Element 5d (Ethical Standards), as it was in year one of the pilot. Element 5a (School Leadership) is also among the highest rated elements. In contrast, 5b (Professional Contributions) is the lowest rated element on the rubric with 37 percent of teachers receiving a rating below proficient. Similarly, though not as extreme, 14 percent of teachers received a below proficient rating on Element 5c (Student Advocate).

The elements within Standard 5 have weak to moderate correlations ($0.21 < \rho < 0.49$). Each element rating is strongly correlated to the overall standard rating ($0.52 < \rho < 0.72$). Reliability analyses also suggest that the ratings demonstrate high internal consistency (Cronbach's $\alpha = 0.71$).

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



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

In summary, the highest and lowest rated elements on the professional practice rubric (in order) are:

Highest rated elements

Element 5d: High ethical standards.
Element 2a: Predictable classroom learning environment in which each student has a positive, nurturing relationship with caring adults and peers.
Element 5a: Demonstrate leadership in their schools.
Element 4a: Analyze student learning, development, and growth and apply what they learn to improve their practice.
Element 4c: Respond to a complex, dynamic environment.

Lowest rated elements

Element 5b: Contribute knowledge and skills to educational practices and the teaching profession.
Element 3e: Communicate high expectations for all students and plan instruction that helps students develop critical-thinking and problem solving skills.
Element 3h: Use appropriate methods to assess what each student has learned, including formal and informal assessments, and use results to plan further instruction.
Element 3d: Integrate and utilize appropriate available technology to maximize student learning.
Element 1b: Knowledge of student literacy development in reading, writing, speaking and listening.

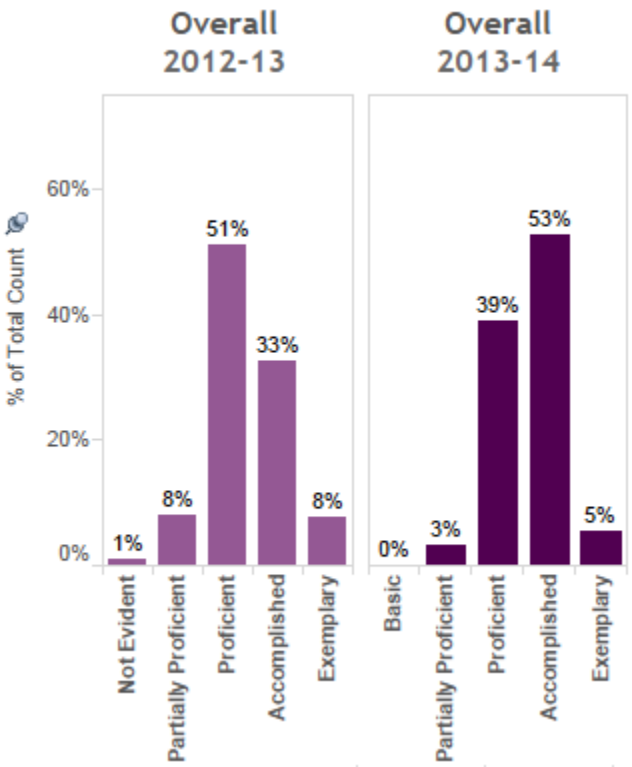
Section 3. Ratings in Subsequent School Year

The foundation of the Colorado State Model Evaluation System for Teachers is the belief that meaningful and specific feedback throughout the school year can improve practice in the short- and long-term. Under an aligned and supportive system, teachers have the opportunity to show professional growth in the current year and subsequent years.

To examine professional growth from year to year, CDE looked at the percentage of teachers who maintained or improved their overall professional practice rating from year one (2012-2013) to year two (2013-2014) as well as the correlations between the two years' ratings. Considering the 1,437 teachers for whom CDE received ratings in both years, 54% of teachers received the same summative rating and 35% improved their performance in year two. Eleven percent of teachers received a lower rating in the second year. The two years' ratings are moderately correlated (ρ = .49).

Looking at the aggregate, there is an upward shift in the distributions from year one to year two with fewer teachers receiving an overall professional practice rating of proficient and more teachers receiving a rating of accomplished (shown in Figure 8). CDE explored whether the teachers who had been in the pilot two years (rather than just one) were driving the upward trend in the distribution. Although teachers for whom CDE had two years of data did have higher rates of accomplished ratings (58 percent were rated accomplished), accomplished was the most common rating for teachers with one year of data as well (49 percent were rated accomplished).

Figure 8. ***Ratings distributions in 2012-2013 and 2013-2014***

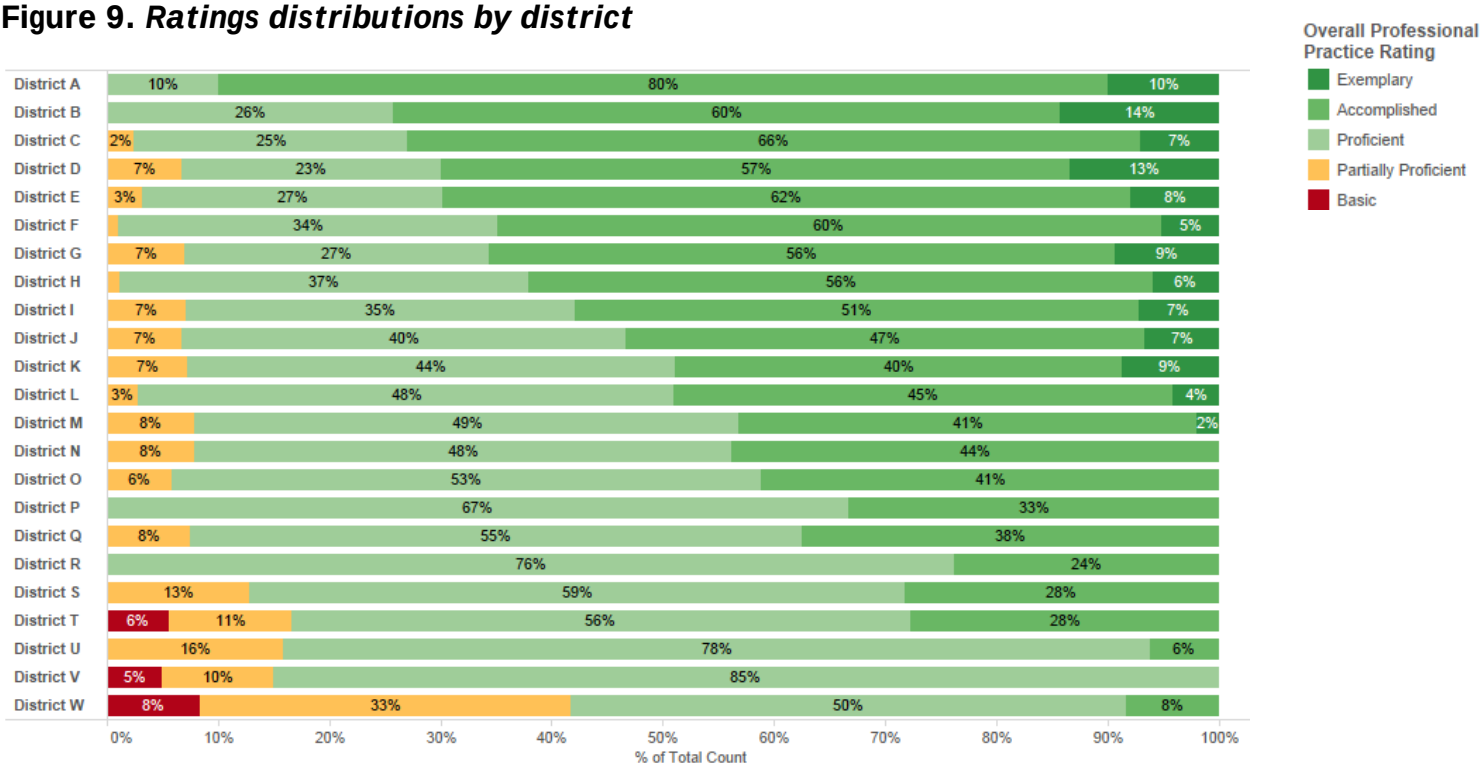


Section 4. Ratings Distributions by District

As stated previously, 25 districts piloted the teacher model evaluation system and 23 of those districts submitted evaluation ratings for the districts' teachers. Differences between the distributions of overall professional practice ratings 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 five teachers participating in the pilot have been removed.

Although there are large differences between districts in the percent of teachers in each performance category, the differences are not as extreme as they were in year one. However, districts' rank order in year one and two (i.e., where they fall in the list of District A through V) is strongly correlated ($\rho = .62$), meaning that districts that had higher ratings in 2012-2013 were more likely to have higher ratings in 2013-2014 as well. For example, the district identified as District A in the graph below was also District A in the prior year. The district identified as District L in the graph below was District I in the prior year, indicating that where individual districts fall in the distribution of all districts is correlated from year to year.

Figure 9. ***Ratings distributions by district***



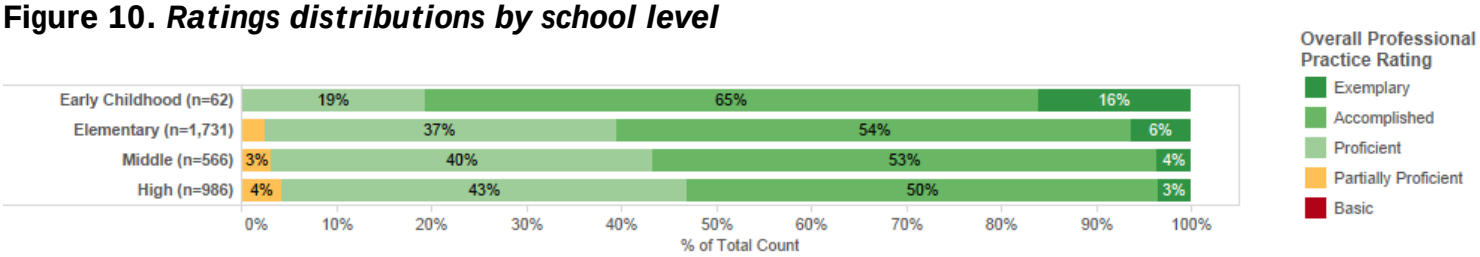
Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by district, meaning that the distribution of overall professional practice ratings varies as a function of the district.

Section 5. Ratings Distributions by Teacher Employment and Demographic Characteristics

This section examines differences in performance category by teachers' employment and demographic characteristics. Figures 10-17 present overall professional practice ratings based on a range of characteristics.[3]

Starting with teachers' employment characteristics, Figure 10 displays the differences between early childhood education (ECE), elementary, middle, and high school teachers. There are statistically significant differences in teachers' overall professional practice ratings based on the school level. ECE teachers received the highest ratings, followed by elementary, middle school, and high school teachers (all group differences are statistically significant except for the difference between middle and high school teachers). The differences based on school level are analogous to the differences CDE found in pilot year one.

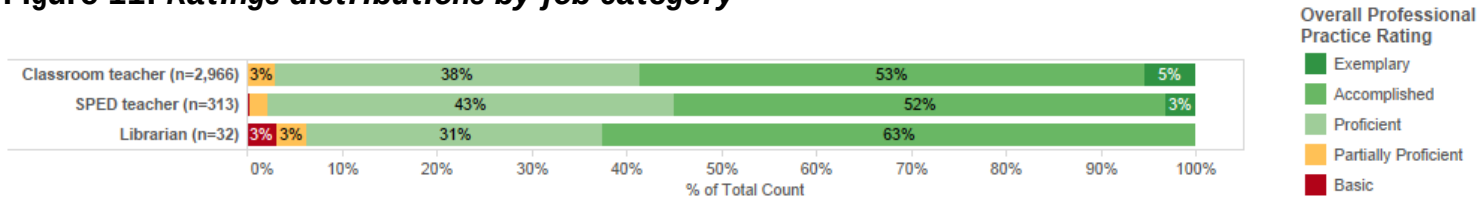
Figure 10. ***Ratings distributions by school level***



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 professional practice 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). The order of ratings is the exact opposite of year one in which librarians received the highest overall professional practice ratings and classroom teachers received the lowest ratings. However, these group differences were not statistically significant in year one or year two.

Figure 11. ***Ratings distributions by job category***

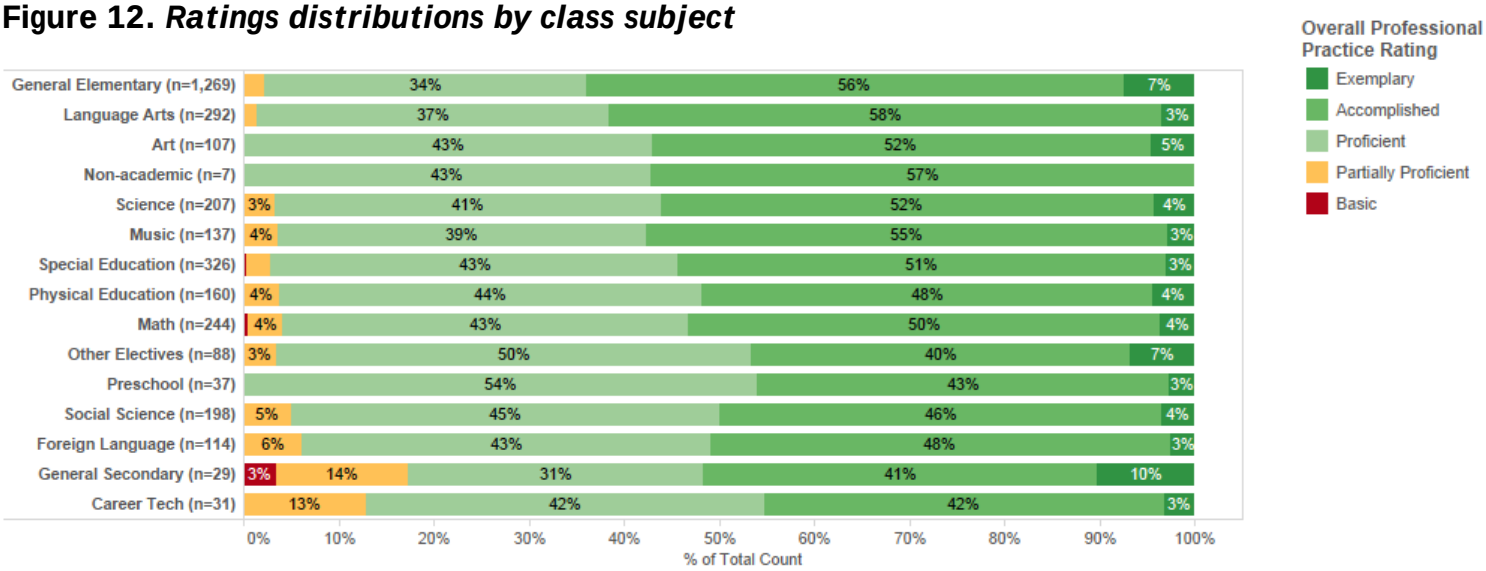


Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are no statistically significant group differences by job category.

[3] Note that the findings exclude groups with fewer than five teachers, CDE does not always have characteristic data for every teacher in the pilot sample, and the characteristic data are from multiple data sources so the n sizes do not always align perfectly.

Limiting the analyses to classroom teachers, there are statistically significant differences between teachers' ratings based on the subject they teach (depicted in Figure 12).[4] The differences indicate that general elementary and Language Arts teachers receive the highest ratings, which is consistent with findings from year one of the pilot.

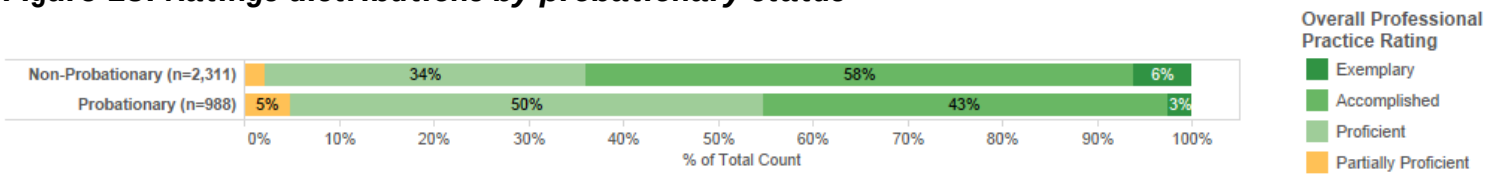
Figure 12. *Ratings distributions by class subject*



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 professional practice ratings varies as a function of the class subject.

The differences in teachers' overall professional practice ratings based on probationary status are statistically significant. Teachers with non-probationary status receive higher ratings (depicted in Figure 13), as they did in year one.

Figure 13. *Ratings distributions by probationary status*

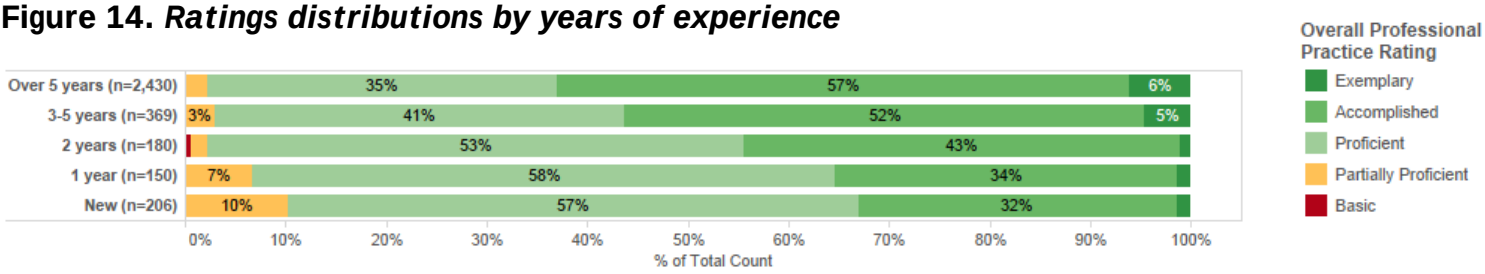


Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by probationary status, meaning that the distribution of overall professional practice ratings varies as a function of teachers' probationary status.

[4] Note that the numbers in Figure 12 (*Ratings distributions by class subject*) do not correspond perfectly to the numbers in Figure 10 (*Ratings distributions by school level*) because the values come from two different data sources.

There also are statistically significant differences in teachers' overall professional practice ratings based on their years of experience. Similar to year one, teachers with more years of experience receive higher ratings.

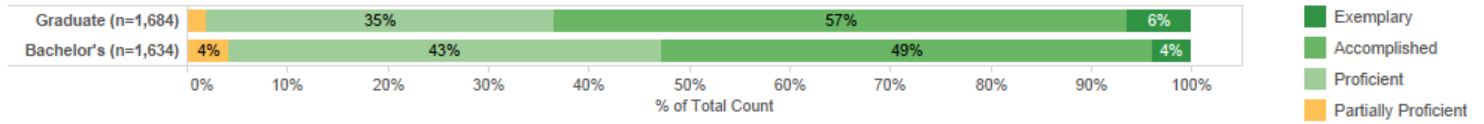
Figure 14. *Ratings distributions by years of experience*



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

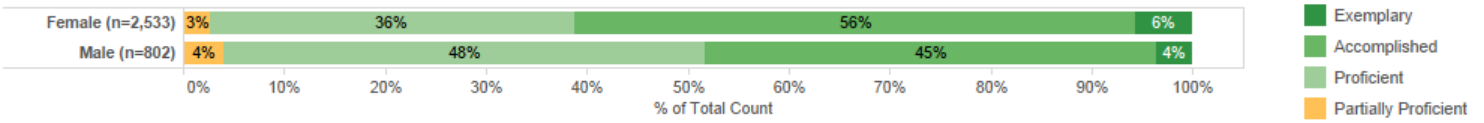
The final group differences in this section pertain to teachers' education level, gender, and race/ethnicity. Teachers with a graduate degree receive higher ratings than teachers with a bachelor's degree (see Figure 15). Females receive higher ratings than males (see Figure 16) and white teachers receive higher ratings than Latino and black teachers (see Figure 17). With the exception of the difference between white and black teachers, all of these group differences were statistically significant in year one as well.

Figure 15. *Ratings distributions based on teachers' education level*



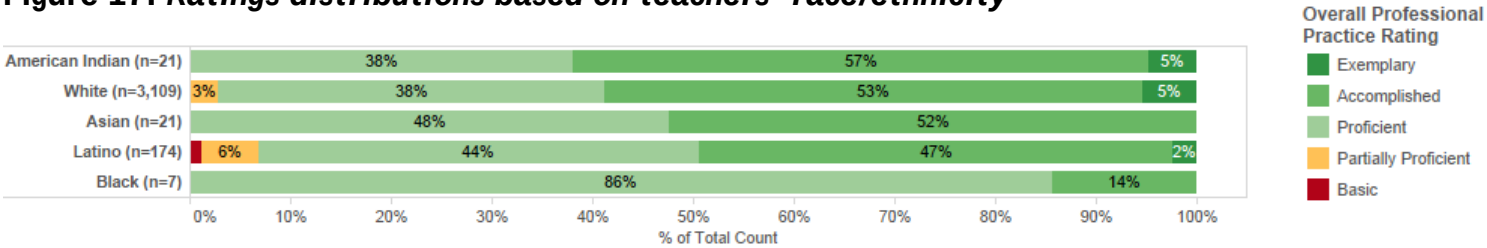
Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by education level, meaning that the distribution of overall professional practice ratings varies as a function of education level.

Figure 16. *Ratings distributions based on teachers' gender*



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

Figure 17. *Ratings distributions based on teachers' race/ethnicity*



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

Section 6. Ratings Distributions by Evaluator Employment and Demographic Characteristics

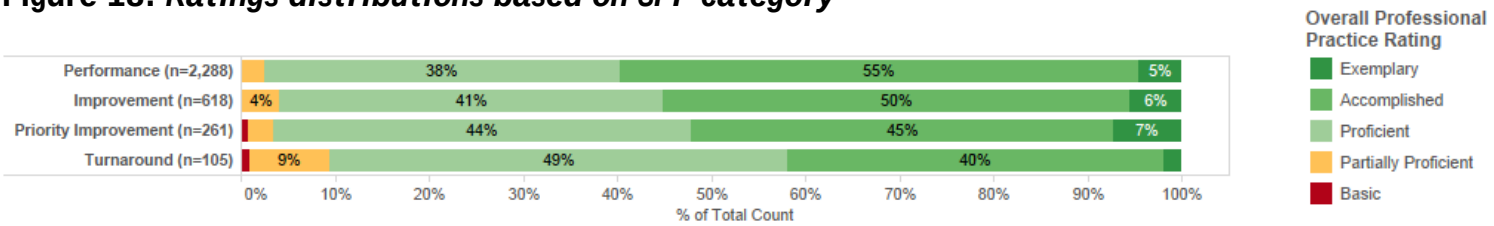
The intent was to report on any group differences based on characteristics of the evaluators in the pilot of the teacher model evaluation system, however CDE found no statistically significant group differences. The group of 212 evaluators is comprised of 119 principals, 60 assistant principals, 5 instructional coordinators, 4 school directors, 3 peer teachers, and 3 superintendents (18 evaluators have undetermined roles). To examine possible group differences, CDE analyzed the ratings assigned based on evaluator role, years of experience (as a principal in the current school, as a principal in any school, and as a teacher), education level, gender, and race. No statistically significant group differences were found for any of the groups, perhaps because of the small sample size.

Section 7. Ratings Distributions by School Characteristics

The final step in the pilot analyses was to look at group differences based on characteristics of the schools in which teachers worked. These school characteristics include the schools' rating on Colorado's school accountability measure (the School Performance Framework; SPF) and student population measures. Measures pertaining to the student population in the school include the extent to which the school is comprised of minority students and students who receive free- or reduced-price lunch (FRL), as well as average student achievement and growth in the school.

Figure 18 reports on differences in overall professional practice ratings based on SPF category (note that these are the preliminary 2014 SPF ratings). There are statistically significant group differences based on SPF category whereby teachers in Performance, Improvement, and Priority Improvement schools all received higher ratings than teachers in Turnaround schools.

Figure 18. ***Ratings distributions based on SPF category***



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

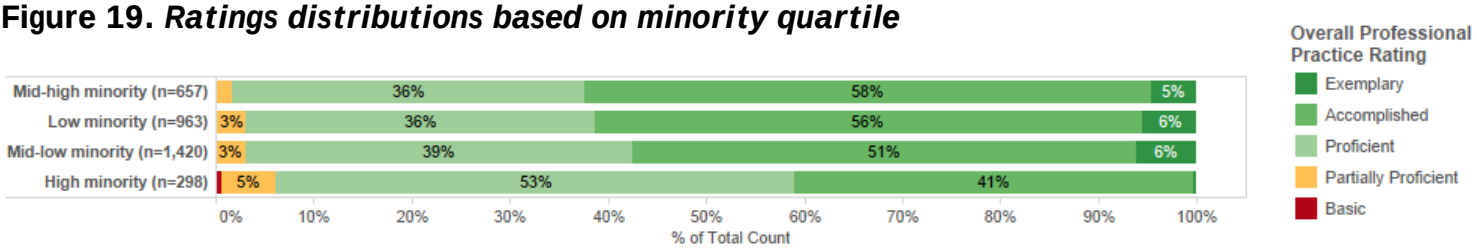
There are also differences based on the demographics of students in the school. Figure 19 displays the differences based on minority quartile and Figure 20 displays the differences based on students qualifying for free- and reduced-price lunch (FRL) quartile.[5] Considering the populations of minority students in schools, the first thing to examine is the

[5] Quartiles are developed at the state level by starting with the percent of minority students in each school, based on the 2013 Student October Count. Once each school has a percentage, the schools are ordered from highest percent of minority students to lowest percent of minority students. Next, four groups of equal size (i.e., equal numbers of schools) are created; these groups are called quartiles. The result at the state level is that 25 percent of schools are in the low minority quartile, 25 percent of schools are in the mid-low minority quartile, 25 percent of schools are in the mid-high minority quartile, and 25 percent are in the high minority quartile. This categorization approach enables the examination of particularly high and particularly low representation of certain groups of students in schools.

representation of low and high minority schools in the pilot sample. In the state of Colorado, the low minority quartile and the high minority quartile are both comprised of 25 percent of schools. In the pilot sample, 33 percent of the schools are from the low minority quartile and 9 percent of schools are from the high minority quartile. These figures indicate that there is a lower percentage of high minority schools represented in the pilot sample than are in the state of Colorado.

As depicted in Figure 19, CDE found statistically significant group differences with teachers in high minority schools receiving lower overall professional practice ratings than teachers in mid-high minority schools, low minority schools, and mid-low minority schools.

Figure 19. *Ratings distributions based on minority quartile*

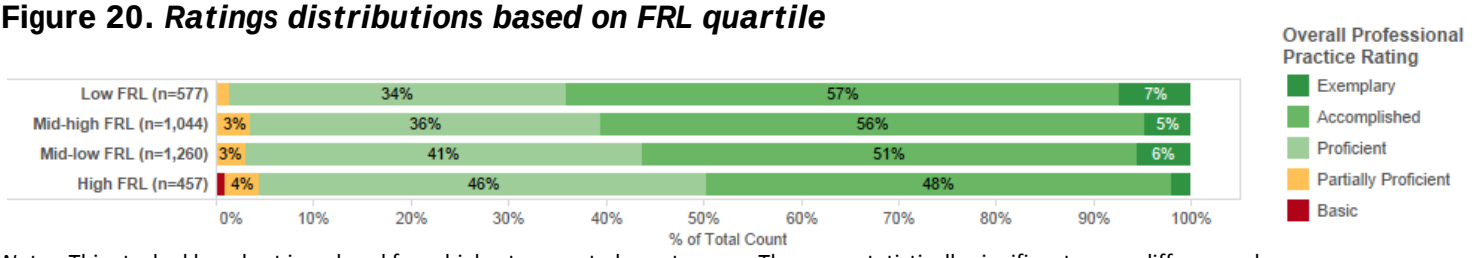


Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by the school's minority quartile, meaning that the distribution of overall professional practice ratings varies as a function of the amount of minority students in the school.

With regard to students qualifying for free- and reduced-price lunch (FRL), in the pilot sample, 17 percent of schools are from the low FRL quartile and 15 percent are from the high FRL quartile. These figures indicate that there are fewer low and high FRL schools represented in the pilot sample than are in the state of Colorado, indicating that there are more schools represented in the mid-low and mid-high FRL quartiles.

When looking at ratings differences based on the amount of FRL students in the school, teachers in schools with more FRL students (defined as high FRL schools) received the lowest overall professional practice ratings (depicted in Figure 20). Their ratings were statistically significantly lower than teachers in low FRL, mid-high FRL, and mid-low FRL schools. Additionally, the difference between teachers' ratings in low FRL schools and mid-low FRL schools was also statistically significant.

Figure 20. *Ratings distributions based on FRL quartile*



Notes. This stacked bar chart is ordered from highest scores to lowest scores. There are statistically significant group differences by the school's FRL quartile, meaning that the distribution of overall professional practice ratings varies as a function of the amount of FRL students in the school.

CDE also examined the aggregated student achievement and growth levels in the school, in this case looking at the average of these student outcomes for teachers in each performance category. For student achievement, CDE examined the average percent of proficient or advanced students on TCAP reading in 2014 for each teacher performance category (reported in Figure 21). For student growth, CDE looked at the average student growth percentile on TCAP reading in 2014 by teacher performance category (reported in Figure 22). Note that averages for the basic category are not displayed because this performance category has fewer than five teachers. The number of teachers represented by the remaining four performance categories corresponds to the numbers and percentages in the Overall Professional Practice graph in Figure 2 (meaning there are larger numbers of teachers in the proficient and accomplished categories).

Figure 21. ***Ratings averages and deviations based on student achievement***

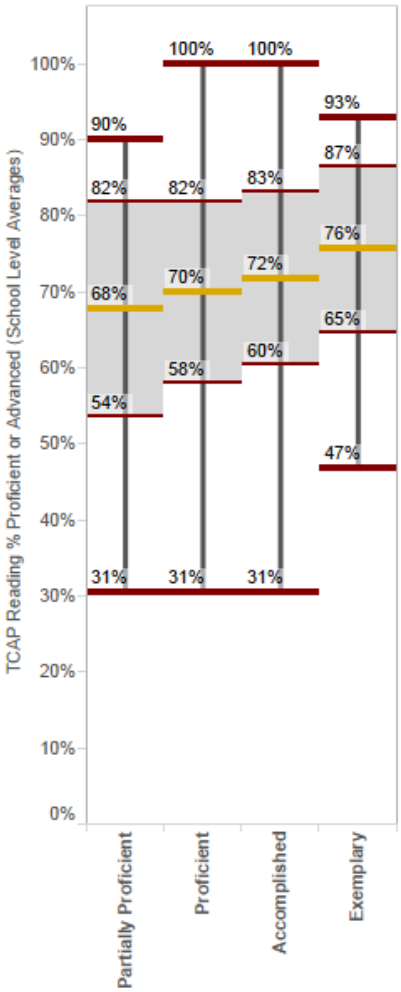
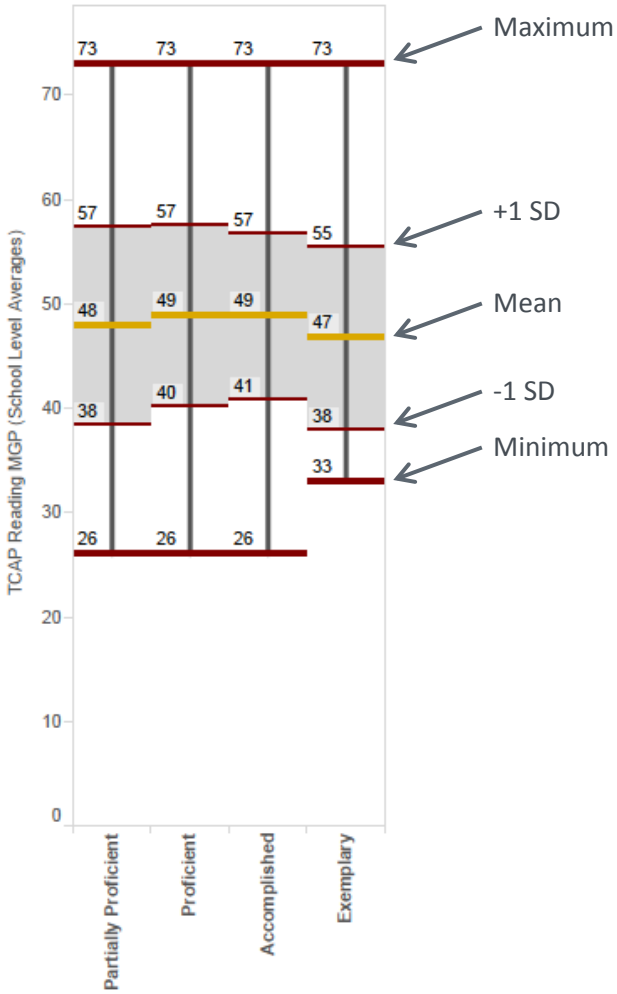


Figure 22. ***Ratings averages and deviations based on student growth***



Looking at the specific information portrayed in the graphs, the yellow line in the middle depicts the mean for that group of teachers and the grey shaded area depicts plus and minus one standard deviation from the mean (also indicated by the thin red line). The interpretation of the area that represents one standard deviation in each direction around the mean is that 68 percent of the teachers in that performance category are captured in that grey band; it is an indication of the spread of scores around the mean. Finally, the thick red lines at the top and bottom of the graph indicate the maximum and minimum in that performance category.

Figure 21 displays findings related to teachers' overall professional practice ratings and the average level of student achievement in the school (note that findings are only reported for TCAP reading but were similar for TCAP math). Considering teachers who received a summative rating of proficient, on average, 70% of the students in their schools scored proficient or advanced on TCAP reading. The average level of student achievement for two-thirds of teachers in the proficient category (specifically, for the 68 percent of the teachers represented within plus and minus one standard deviation) falls between 58 and 82 percent of students proficient or advanced on reading. But some teachers do work in schools with far lower and higher achievement. The minimum achievement level in this category is 31 percent, meaning that some teachers who receive a summative rating of proficient work in schools where 31 percent of students are proficient or advanced in reading. Conversely, other proficient-rated teachers work in schools where 100 percent of the students are proficient or advanced in reading. These figures can be compared to the averages and ranges for the other teacher performance categories.

Figure 21 also displays the general pattern of the relationship between teachers' overall professional practice ratings and the percent of proficient and advanced students in their school. In general, as teachers' summative ratings increase the achievement level of the students in the school does as well. The correlation between these two measures is statistically significant but weak ($\rho = .11$).

Figure 22 summarizes similar analyses but pertaining to student growth instead of student achievement. Here CDE reports on the relationship between teachers' overall professional practice ratings and the average level of student growth on TCAP reading in the school, aggregated up as a mean growth percentile (MGP).[6] For teachers who received a summative rating of proficient, on average the students in their schools have a reading mean growth percentile of 49. Two-thirds (68 percent) of the teachers work in schools with a reading MGP between 41 and 57. The lowest school-wide MGP for proficient teachers is 26 and the highest school-wide MGP is 73.[7] There is no general pattern of the relationship between teachers' overall professional practice ratings and student growth in their school; this relationship is not statistically significant.

[6] The mean was selected instead of the median because utilizing the mean enables an examination of variance around the mean. Medians are another measure of an "average" but are based on rank order and therefore do not have variance around the average. Without variance it is impossible to examine the standard deviation, and the standard deviation is useful because it gives information about how spread out numbers are in a distribution. A recent report on Using Student Growth Percentiles for Educator Evaluations (and the executive summary) discusses the technical consideration of using a mean versus a median.

[7] It should be noted that the figures change slightly if examining the *median* growth percentile instead of the *mean* growth percentile. The median value is 50 for the partially proficient category (higher than the mean), 48 for the proficient category (lower than the mean), 49 for the accomplished category (same value), and 45 for the exemplary category (lower than the mean). Essentially, medians are more susceptible to outliers in a small sample size so the medians of the partially proficient and exemplary categories are pulled further out (higher for partially proficient and lower for exemplary).

Summary and Next Steps

As with year one of the pilot of the Colorado State Model Evaluation System for Teachers, year two findings indicate that the professional practice rubric captures multiple aspects of teaching and differences in teacher practice. Relatedly, evaluators continue to make meaningful distinctions between teachers and with regard to different aspects of practice. Educators receive the highest ratings on elements related to reflection on their practice and the lowest ratings on elements related to effective instruction and facilitating learning for all students.

CDE continues to find evidence of growth in teachers' practice during the school year. In year one, the majority of teachers maintained or improved their practice through the course of the school year. Similarly, the majority of teachers maintained or improved their practice in the following school year. The model evaluation system is built on the belief that teachers who receive clear and frequent feedback on their teaching will improve their practice and therefore further impact student learning. Both years of pilot findings provide evidence for growth in teacher practice. Additional (forthcoming) analyses are needed to examine the impact of teacher practice on student learning and how teachers are feeling about the process and the usefulness of the feedback they are receiving.

CDE found evidence of group differences based on a variety of district, school, teacher, and student characteristics. It is important to remember that although these factors are associated with teachers' overall professional practice ratings, it does not indicate that one is causing the other (e.g., being an ECE teacher necessarily causes one to earn higher ratings). Interpretations for these findings range from a true reflection of teacher skill within certain groups, evaluator comprehension of the rubric's application to a range of teacher types, evidence for additional skills needed as evaluators, a product of the text in the professional practice rubric, and district policies in evaluation processes as well as a range of other plausible explanations. For these reasons, CDE will explore explanations for the findings but also must interpret cautiously and avoid making causal connections.

Pilot analyses will continue on the 2013-2014 teacher model evaluation system data, particularly focusing on analyses related to Standard 6, or measures of student learning (MSL). Slight changes were already made to the teacher rubric for the 2014-2015 school year based on the findings from this analysis and feedback from the field, including streamlining portions of the rubric. These changes and others will impact future findings so CDE will continue to collect and analyze data from pilot districts through the 2015-2016 school year. CDE will also be able to provide a statewide analysis after the spring of 2015 as a result of all districts in the state submitting Standards 1-6 and overall ratings (performance ratings for the 2013-2014 school year are submitted through the 2014-2015 HR collection). These analyses will help CDE determine if data from the pilots is representative of the state as a whole.