REVISED STATE MODEL EVALUATION SYSTEM FOR TEACHERS PILOT RESULTS

COLORADO
Department of Education

OVERVIEW

Senate Bill 10-191, passed in 2010, changed the way teachers are supported and evaluated in Colorado. The ultimate goal of the change was to impact the effectiveness of teachers and leaders in schools, which in turn helps to ensure college and career readiness for all students. To support this effort, the Colorado Department of Education (CDE) developed a model system as an option for districts and BOCES to use in implementing the new evaluation requirements for educators. During the 2017-18 school year, 83% (163 of 196) local education agencies (LEAs) that employ teachers opted to utilize the Colorado State Model Evaluation System (SMES) for teachers.

As part of the Department's commitment to continuous improvement, CDE obtained broad stakeholder input regarding suggested improvements and revisions to the SMES for teachers. Several adjustments to SMES for teachers were made, including a reduction in the number of Teacher Quality Standards included on the professional practices rubric from 5 to 4; a reduction in the number of Elements included within each Standard from 27 to 17; a reduction in the number of professional practices identified on the rubric; category label changes on the rubric to promote more constructive conversations with teachers about their practices; a shift to a preponderance of evidence requirement for scoring at the standard level such that a majority of elements must be rated at a higher level for the standard to be rated at the higher level; and the requirement that at least the midpoint score in the Accomplished level is achieved by an educator to be eligible for an overall Highly Effective rating.

To pilot test these revisions, CDE solicited voluntary participation from LEAs for the 2017-2018 school year. Fifty districts and BOCES in Colorado, reflecting approximately one-quarter of all districts and BOCES in the state, opted into the pilot program to investigate whether these adjustments seemed appropriate for statewide adoption.

The current report presents findings from the 2017-2018 pilot of the revisions to the rubric for and scoring of professional practices in the State Model Evaluation System for teachers. Findings are based on multiple sources of evidence, which include evaluation ratings, perception survey results, and focus group themes. More specifically, this report draws upon evaluation ratings for 14,469 teachers representing all 50 participating pilot districts and BOCES; perception survey results representing all 50 pilot districts and BOCES from 5341 teachers, 826 principals, and 62 administrators on the pre-survey administered in September 2017, and 4279 teachers, 379 principals, and 48 administrators on the post-survey administered in April 2018; and themes from 19 focus groups conducted in April 2018 that reflect input gathered from 82 teachers, 19 principals, and 17 district leaders.

Data reported here are based exclusively on these 50 LEAs and are not statewide results. Results reported from this pilot cannot be assumed to extrapolate necessarily to statewide results because the participating LEAs were voluntary. However, results are reported in comparison to these same districts and BOCES in the prior year, wherever possible, to provide a sharper focus on their experiences.

Key Findings

The revisions to the SMES for teachers were designed to result in specific improvements that educators and other stakeholders requested, and results from this year-long pilot of the revised system indicate that the intended outcomes were achieved.

- Comparison of 2016-17 *overall educator effectiveness ratings* to 2017-18 *overall educator effectiveness ratings* in the pilot sites shows that, as anticipated and consistent with the refined rubric with increased performance expectations and increased scoring rigor in the pilot test, the overall educator effectiveness evaluation ratings

are distributed differently in the pilot year than they were in the prior year. In the pilot year there were similar percentages of teachers rated Effective or above, 75.6%, compared to the prior year, 77.6%. However, the pilot data had substantially fewer Highly Effective teachers, 15.6% compared to 35.9% (a reduction of 20.3%), and more Effective teachers, 60.0% compared to 41.7% (an increase of 18.3%). The decrease in the Highly Effective category essentially off-set the increase in the Effective category, resulting in a 2% decrease in teachers rated Effective or above. While the rubric and scoring changes may have addressed concerns about an inflated number of teachers in the highest category, the changes seem to have had little impact on the representation of teachers in the lower two performance categories.

- Similar to the overall effectiveness ratings, the distribution of teacher ratings on their *professional practices* was different in 2017-2018. The more rigorous rubric and preponderance of evidence scoring approach resulted in more teachers earning *professional practice ratings* in lower performance levels. The largest shifts are within the three categories that indicate meeting state teaching standards at a proficient or higher performance level. More specifically, in 2017-18 many more teachers' practices are rated Proficient (49.5% compared to 37.7%), fewer are rated Accomplished (29.9% compared to 51.0%) and fewer are rated Exemplary (1.8% compared to 7.7%) than in the previous year. The shift in the categories below Proficient is smaller (a combined 10.4% in 2017-18 compared to 3.1% in 2016-17, or 7.3% more).
- Overall evaluation ratings vary by certain teacher demographics, job classification, and content area taught, as was evident in earlier studies of the SMES. However, it is important to note that there are a variety of possible explanations and contributing factors, ranging from true reflection of teacher skill, variation in evaluator training and skill, quality of Measures of Student Learning variation, district policies regarding evaluation processes, as well as many other influences.
- There was a strong consensus that streamlining the teacher rubric was a beneficial change leading to a more efficient use of evaluation time and a more authentic evaluation process.
    - Educators specifically mentioned having more time in the evaluation process for shared understandings and more focused dialogue.
    - Principals reported that using the new rubric enabled more coherent goal setting because they were more able to review self-assessments prior to observations, identify trends, and engage in more meaningful conversations as part of the evaluation process.
- There was strong support for the new rubric's combining elements of professional reflection and leadership into a single standard, *Quality Standard IV – Teachers demonstrate professionalism through ethical conduct, reflection, and leadership*.
- Multiple teachers, principal/evaluators, and administrators expressed appreciation for the scoring adjustments that they expected to shift some teachers from the highly effective category to the effective category. There was some concern about difficult conversations that might arise as a result, although some pointed out that difficult conversations can have positive results.
- Survey and focus group feedback also provided direction to CDE for the next cycle of continuous improvement. One area that surfaced as a persistent concern across focus groups relates to local capacity for implementation of the evaluation system. This arose in conversations about rater reliability, time burdens, and capitalizing on evaluation information to inform professional development offerings, among others. As one way to ameliorate local capacity challenges, many pilot educators made recommendations for targeted areas of support that might be offered to the field by the Educator Effectiveness staff. Recommendations included "over-communication" from CDE to the field regarding the fundamental purposes of the teacher evaluation process and the rubric changes and dissemination of an updated resource guide aligned to the new rubric, along with suggestions for specific actions that could be taken locally to strengthen teacher ownership during the evaluation process. Multiple respondents also recommended local leeway to tailor evaluation frequency based on teachers' previous evaluation rating, although there was recognition that this is inconsistent with current law.

CONTENTS

SUMMARY OF SYSTEM CHANGES IMPLEMENTED IN THE PILOT 3
DATA COLLECTION AND RESULTS 4
DISTRIBUTION OF OVERALL EDUCATOR EFFECTIVENESS 5
DISTRIBUTION OF PROFESSIONAL PRACTICES 6
DISTRIBUTION OF TEACHER QUALITY STANDARDS 6
DISTRIBUTIONS OF ELEMENTS 9
DISTRIBUTIONS OF SCORES BY TEACHER DEMOGRAPHIC GROUP 11
DISTRIBUTIONS OF SCORES BY TEACHER EMPLOYMENT CATEGORIES 13
EDUCATOR SURVEY RESPONSES 19
TEACHER SURVEY RESPONSES 21
PRINCIPAL/EVALUATOR SURVEY RESPONSES 23
DISTRICT ADMINISTRATOR SURVEY RESPONSES 25
THEMES FROM THE FOCUS GROUPS 27

SUMMARY OF SYSTEM CHANGES IMPLEMENTED IN THE PILOT

As part of the Department's commitment to continuous improvement, ad hoc feedback from the field over the course of time led to a formal process to solicit input on the State Model Evaluation System (SMES) for teachers. The formal input process included multiple stakeholder groups, including a rubric Technical Working Group, a scoring Technical Working Group, and 30 focus groups conducted throughout the state to gather feedback about what is working and what could be improved. Data collected largely indicated that the rubric's conceptual integrity should be preserved and constructs measured should remain the same, while also calling for streamlining, clarification of language, category relabeling, and more rigorous scoring.

As a result of the formal input process, several aspects of the State Model Evaluation System experienced revision. With this broad base of input to inform the revisions, CDE solicited voluntary participation to pilot test the system changes during the 2017-2018 school year. All changes were piloted by 50 districts and BOCES in Colorado, reflecting approximately one quarter of all districts and BOCES, to investigate whether further adjustments seem warranted prior to statewide release of the revised SMES for teachers.

This report begins with a brief outline of the revisions to the SMES for teachers that were pilot-tested during the 2017-2018 school year; further details, including a list of participants and project timeline, can be found on CDE's website for the State Model Evaluation System Revisions. The 2017-2018 pilot test included the following system revisions:

- Reduction in the number of Teacher Quality Standards included on the professional practices rubric from 5 to 4.
- Reduction in the number of Elements included within each Standard from 27 to 17.
- Reduction in the number of professional practices identified on the rubric; for example, from 309 for an elementary teacher to 167.
- Category label change on the rubric to promote more constructive conversations with teachers about their practices. The revised rubric uses categories of Level 1 Practices, Level 2 Practices, Level 3 Practices, Level 4 Practices, and Level 5 Practices. (Categories remain unchanged for overall professional practices and standards and continue to be Basic, Partially Proficient, Proficient, Accomplished, and Exemplary.)
- Shift to a preponderance of evidence requirement for scoring at the standard level such that a majority of elements must be rated at a higher level for the standard to be rated at the higher level.
- Requirement that at least the midpoint score in the Accomplished level is achieved by an educator to be eligible for an overall Highly Effective rating.

Having explained the rationale and support for SMES adjustments and provided a brief overview of them, we next address the data collected in the pilot sites.

DATA COLLECTION AND RESULTS

This section begins the discussion of the pilot test data collected and implications drawn. In an effort to collect data on several aspects of the system and from as many stakeholders as possible, the Department implemented several data collection activities involving the 50 sites, including:

- Obtaining all 2017-2018 teacher evaluation ratings, inclusive of overall ratings, professional practices ratings, standards I through IV ratings[1], and element ratings;
- Obtaining all 2016-2017 teacher evaluation ratings, inclusive of overall ratings, professional practices ratings, standards I through V ratings[2], and element ratings;
- Administering pre-pilot perception surveys to teachers, principals and other evaluators, and district administrators based on experiences with the SMES during the 2016-2017 year;
- Administering post-pilot perception surveys to teachers, principals and other evaluators, and district administrators based on experiences with the revisions in the 2017-2018 year; and
- Conducting, through collaboration with WestEd, 19 focus groups that occurred across the state in the spring of 2018, reflecting input from teachers, principals, and district leaders[3].

Data analysis included comparison of the pilot sites' evaluation ratings to the prior year, disaggregation of scores by teacher demographics and employment characteristics, correlational relationships among standards, pre/post statistical analysis and effect size analysis of surveys within teacher, principal/evaluator, and administrator groups, thematic analysis of focus group and open-ended survey responses, and triangulation of findings across data sources.

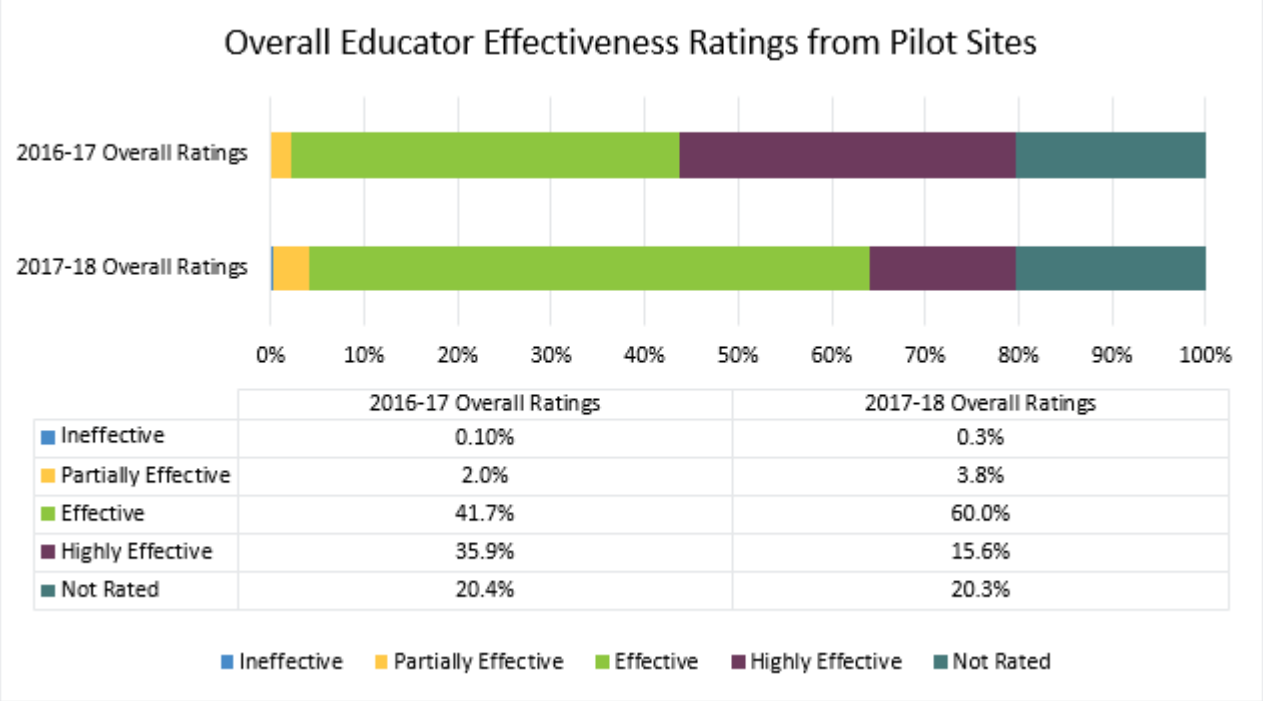
---

[1] Standard V in 2017-2018 and Standard VI in 2016-2017 are based on measures of student learning, which are separate from the professional practices rubric and scoring. There were no changes implemented related to that standard in the pilot, so it is not included in this report.

[2] See footnote 1.

[3] More detail about the focus groups conducted with the support of WestEd in April 2018 is available upon request. Please contact EducatorEffectiveness@cde.state.co.us.

DISTRIBUTION OF OVERALL EDUCATOR EFFECTIVENESS

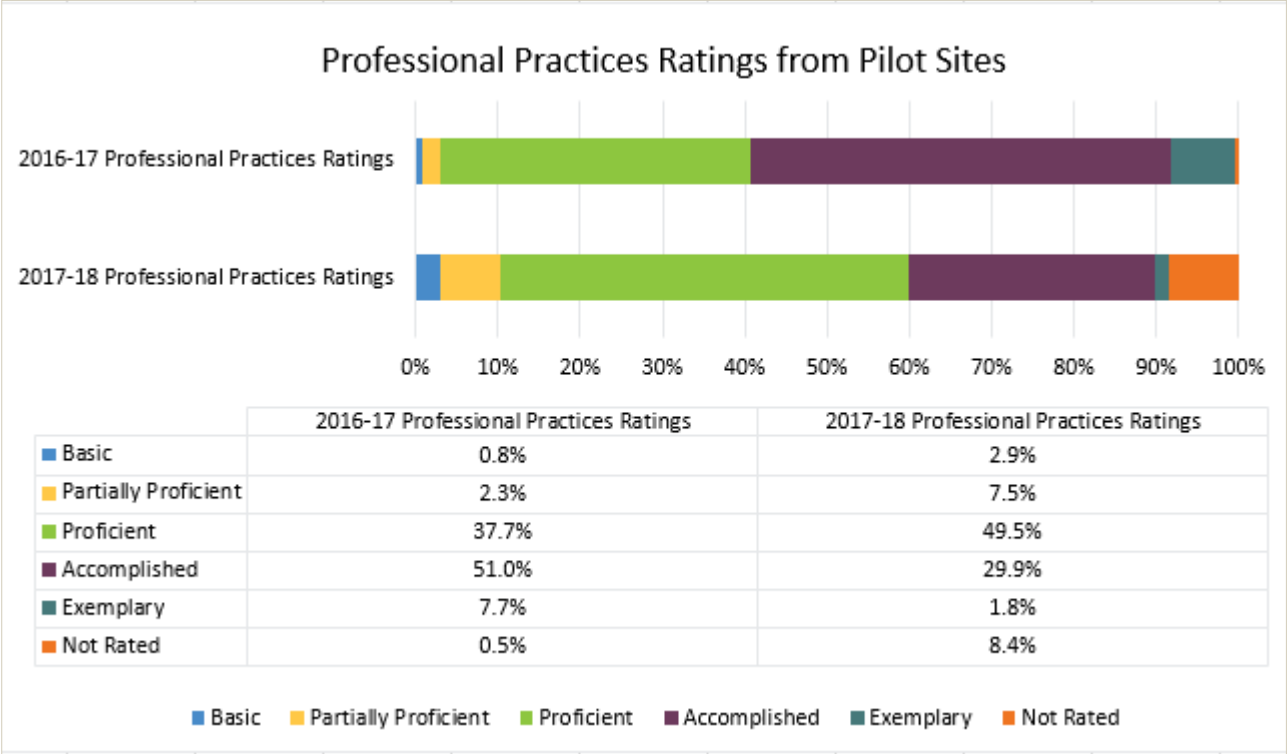


	2016-17 Overall Ratings	2017-18 Overall Ratings
Ineffective	0.10%	0.3%
Partially Effective	2.0%	3.8%
Effective	41.7%	60.0%
Highly Effective	35.9%	15.6%
Not Rated	20.4%	20.3%

As anticipated, and consistent with the more explicit rubric and rigorous scoring in the pilot test, the overall educator effectiveness evaluation ratings are distributed differently in the pilot year than they were in the prior year. In the pilot year, there were similar percentages of teachers rated Effective or above, 75.6%, compared to the prior year, 77.6%. However, the pilot data had substantially fewer Highly Effective teachers, 15.6% compared to 35.9% -- a reduction of 20.3%, and more Effective teachers, 60.0% compared to 41.7% -- an increase of 18.3%. The decrease in the Highly Effective category essentially off-set the increase in the Effective category, resulting in a 2% decrease in teachers rated Effective or above. There were approximately 2% more teachers rated Partially Effective and minimal (less than .5%) change in the Ineffective and Not Rated categories.

While the rubric and scoring changes may have addressed concerns about an inflated number of teacher ratings in the highest category, the changes seem to have had little impact on the representation of teachers in the lower two performance categories.

DISTRIBUTION OF PROFESSIONAL PRACTICES



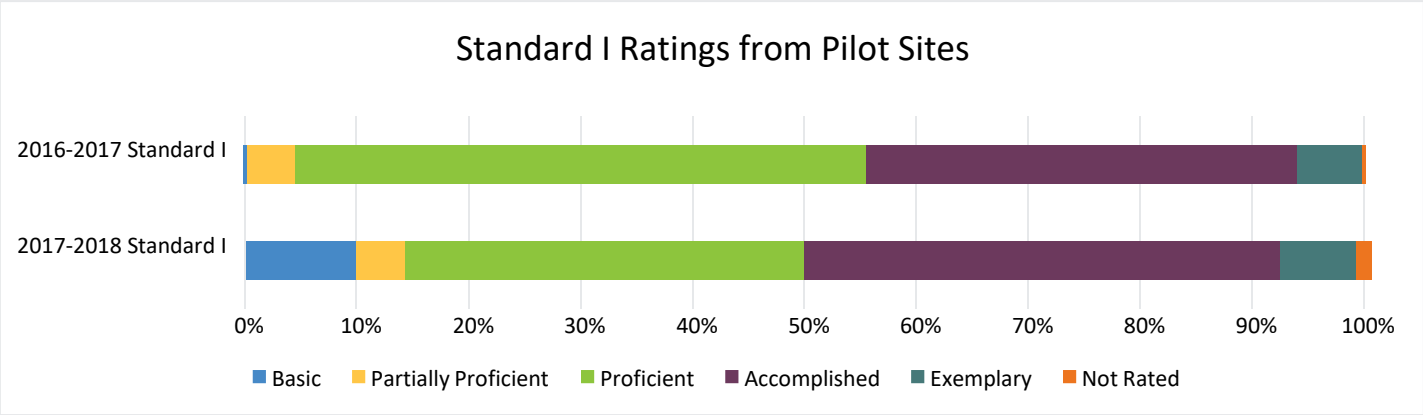
	2016-17 Professional Practices Ratings	2017-18 Professional Practices Ratings
Basic	0.8%	2.9%
Partially Proficient	2.3%	7.5%
Proficient	37.7%	49.5%
Accomplished	51.0%	29.9%
Exemplary	7.7%	1.8%
Not Rated	0.5%	8.4%

Similar to the overall effectiveness ratings, the distribution of teachers' ratings on their professional practices was different in 2017-2018. The largest shifts are within the three categories that indicate meeting state teaching standards at a proficient or higher performance level. More specifically, in 2017-18 many more teachers are rated Proficient (49.5% compared to 37.7%), fewer are rated Accomplished (29.9% compared to 51.0%) and fewer are rated Exemplary (1.8% compared to 7.7%) than in the previous year. The shift in the categories below Proficient is smaller (a combined 10.4% in 2017-18 compared to 3.1% in 2016-17, or 7.3% more). It is important to note that 8.4% of teachers had not yet been awarded a 2017-2018 professional practice rating at the time of this analysis and the impact of this is unknown.

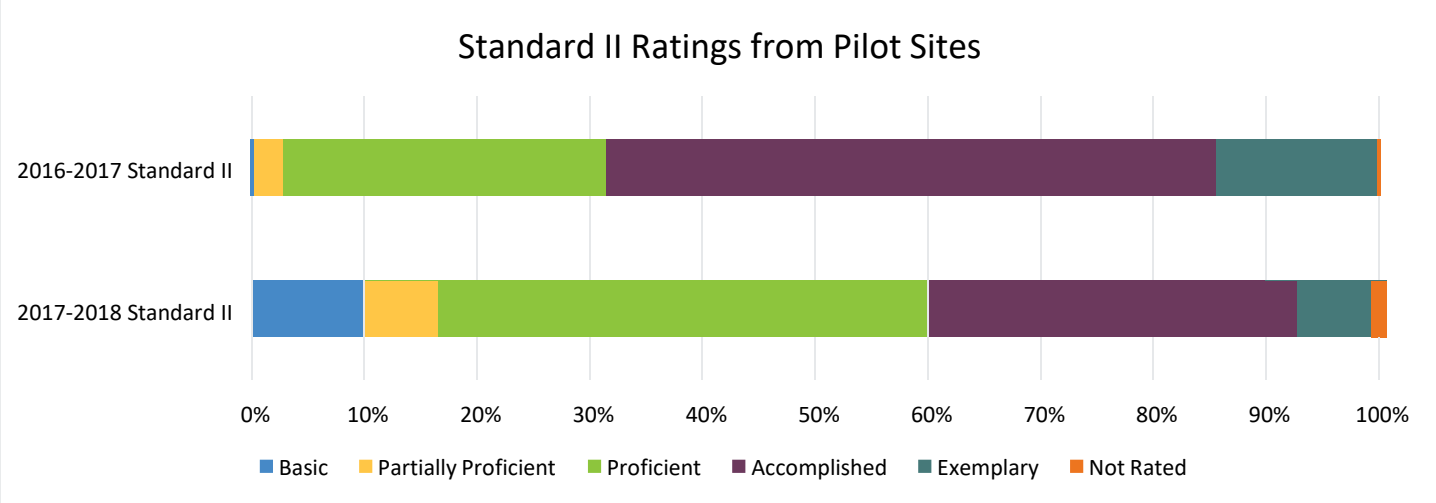
DISTRIBUTION OF TEACHER QUALITY STANDARDS

The distribution of 2017-2018 ratings on the teacher quality standards reflects the increased rigor of the rubric and scoring process, similar to what is observed for the overall effectiveness and professional practice ratings. In general, the distributions of teacher ratings on the standards shifted from being concentrated among the higher levels to being more distributed among all possible levels. As stakeholders desired, the new tool may be better able to discern differences in teacher practice.

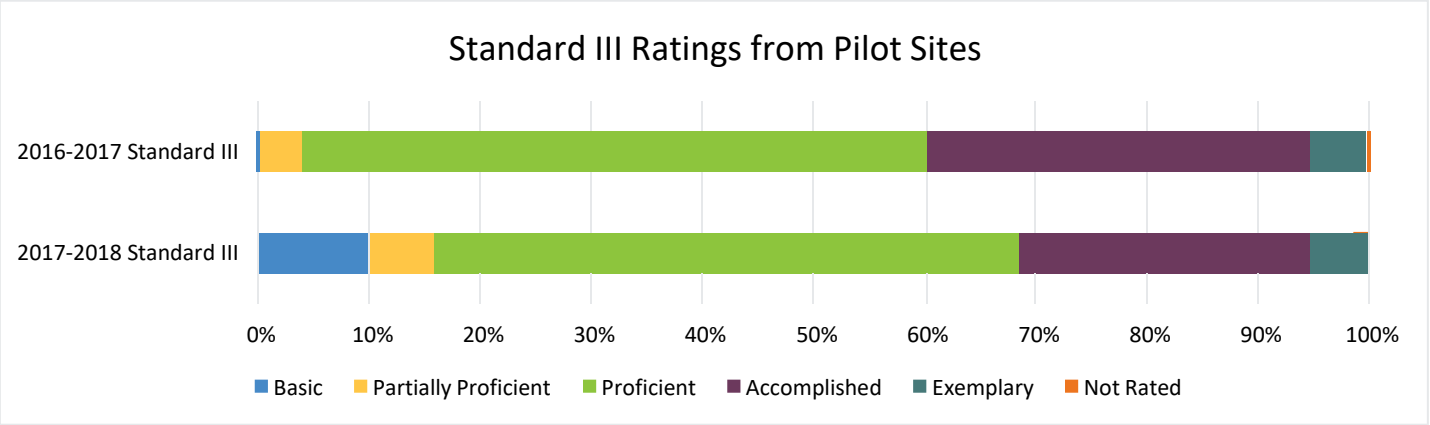
Below, the shifts in distributions for each quality standard are presented in more detail.



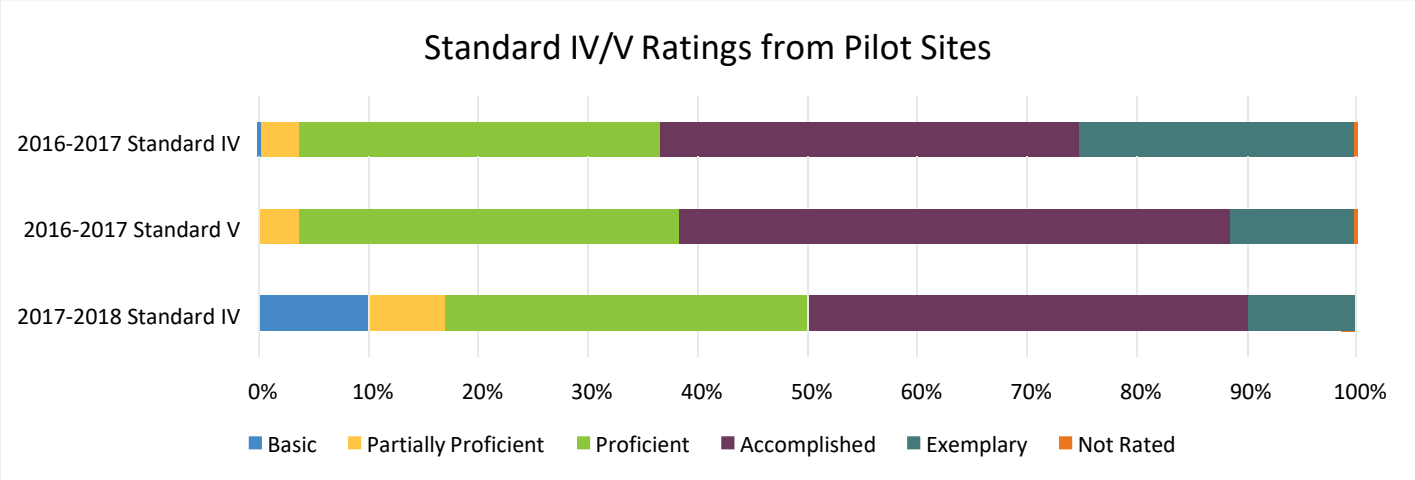
The distribution of ratings on Standard I, demonstrating mastery of and pedagogical expertise in content taught, differed from the prior year. The largest changes were the increase in percentage of ratings in the Basic level (9.5% compared to 0.4%, an increase of 9.1 %) and the decrease of ratings in the Proficient level (35.2% compared to 51.1%, a decrease of 15.9%). Additionally, the percentage of ratings of Accomplished increased by 4.6%. In total, in 2017-18, 84.2% of teacher ratings were Proficient or higher on expertise in the content area taught compared to 95.6% of ratings the prior year.



The distribution of Standard II, establishing a safe, inclusive, and respectful learning environment for a diverse population of students, also exhibits a different pattern in 2017-2018. The total percentage of teachers rated proficient and higher is 81.9% compared to 96.9% the year prior, which, again, meets the objectives of increasing the rigor of both the rubric and scoring as advocated by stakeholders. There are shifts of nearly 10 percentage points or more in all categories except for Partially Proficient, with the largest shifts being a decrease in Accomplished (32.7% compared to 54.2% the prior year, a decrease of 21.5%) and an increase in Proficient (43.4% compared to 28.7%, an increase of 14.7%).



The distribution for Standard III, planning and delivering effective instruction to facilitate student learning, reveals a pattern shift similar to that of the previous standard. Standard III ratings were 82.6% proficient or higher in the pilot year compared to 95.7% proficient or higher the prior year, which, again, reflects the stakeholder-driven goals of the revision, namely a more rigorous rubric and a more stringent approach to scoring. While the Exemplary category remained the about the same (a reduction of 1.1%), the performance levels of Proficient and Accomplished were reduced and the two lower performance levels of Partially Proficient and Basic were increased. The largest changes were in the Accomplished level (a reduction of 7.6%) and the Basic level (an increase of 9.2%).



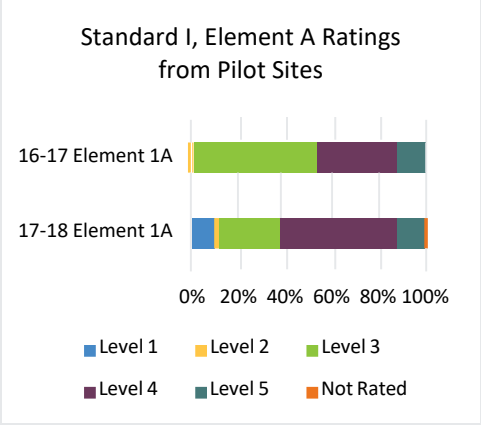
Standard IV in the pilot, demonstrating professionalism through ethical conduct, reflection, and leadership, is a newly constructed standard for this pilot and it is designed to encompass standards IV and V of the original Teacher Quality Standards. As such, the data from Standard IV (Professionalism) for the pilot year is compared here to data for Standard IV (Reflecting on Practice) and Standard V (Demonstrating Leadership) from the prior year. The pattern found in the distributions for the other standards is also found here. There are teacher ratings in the Exemplary category (7.8% in 2017-2018 compared to 24.9% and 11.1% in 2016-2017) and a greater number in the Partially Proficient category (7.2% in 2017-2018 compared to 3.2% and 3.6% in 2016-2017) and Basic category (9.7% in 2017-2018 compared to 0.4% and 0.3% in 2016-2017).

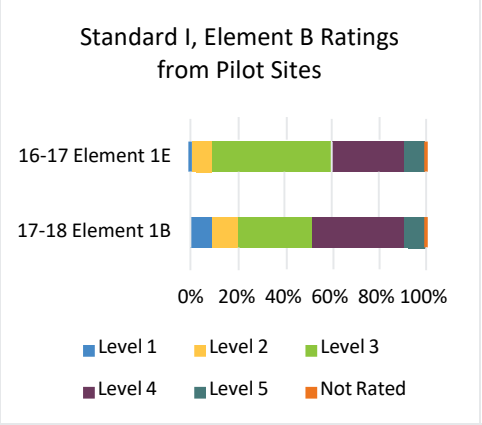
DISTRIBUTIONS OF ELEMENTS

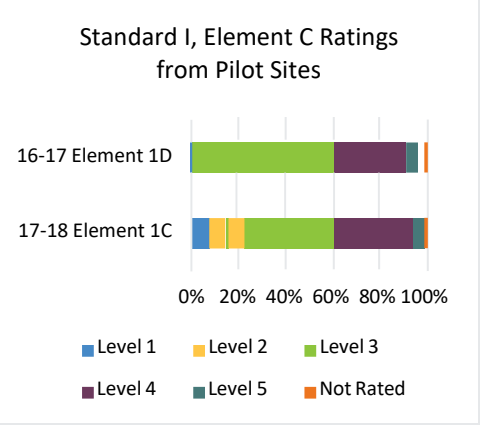
In this next section, distributions of the individual elements that comprise each standard are displayed for both the prior year and the pilot year. The elements within each standard contain more specific professional practices that are necessary for teachers to meet the state standard. The number of elements within each standard varies from three (for Standard I) to six (for Standard III). Elements are scored Level 1 Practices through Level 5 Practices, in contrast to the category labels used for Standards, Professional Practices, and Overall Ratings, to further emphasize a developmental continuum. As mentioned above, category label changes were made on the rubric to promote more constructive conversations with teachers about their practices. The graphs use the Level 1 Practices through Level 5 Practices, although the 2016-17 elements were rated on the Basic, Partially Proficient, Proficient, Accomplished, and Exemplary scale.

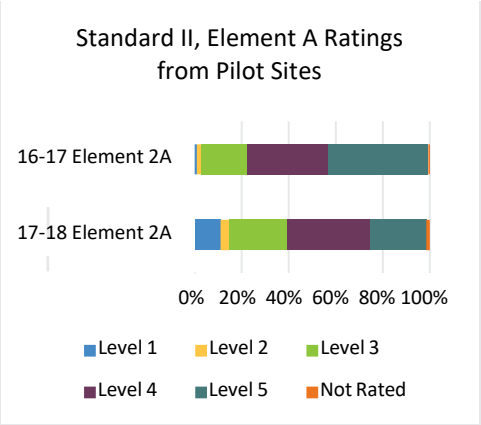
As stated earlier, one of the goals of the pilot was to reduce redundancies and clarify language used in the rubric. As part of the effort to achieve that goal, some of the original elements comprising the TQS were eliminated or combined in the pilot. The charts below present the 2017-2018 elements with the appropriately comparable 2016-2017 elements. (Note that throughout this discussion of the elements, the 2016-2017 data represent from 45 to 48 sites that were in existence the prior year and able to provide prior year data.)

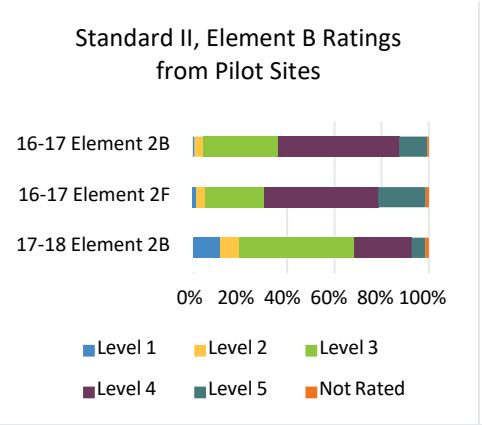
The following graphs are grouped by standard and elements within that standard. In the graphs, elements are presented as Level 1 through Level 5 in place of Level 1 Practices to Level 5 Practices due to space considerations. For the sake of clarity, wherever possible a one-to-one connection was made between the primary 2016-17 element and the 2017-18 element; a one-to-one comparison of element shifts and changes is available on the CDE website.

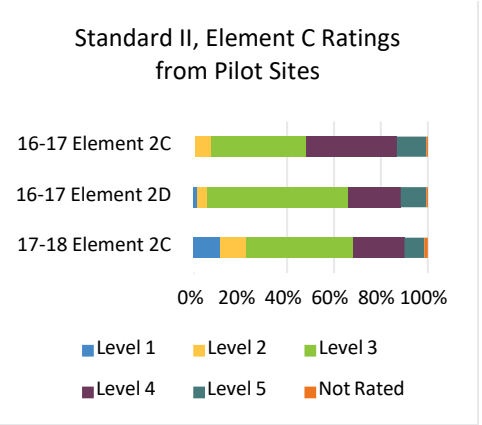


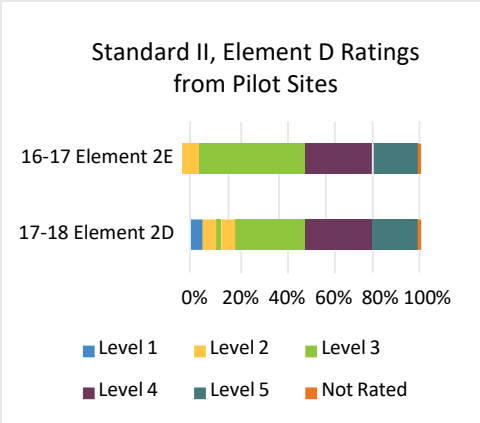


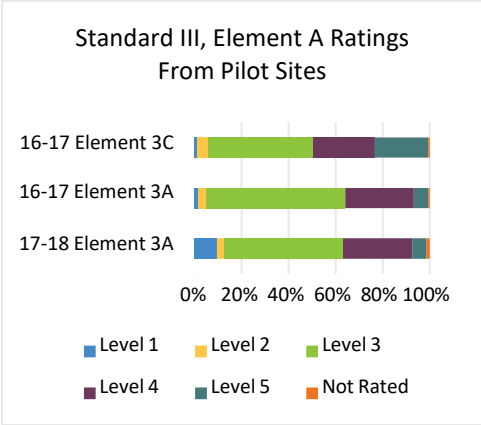


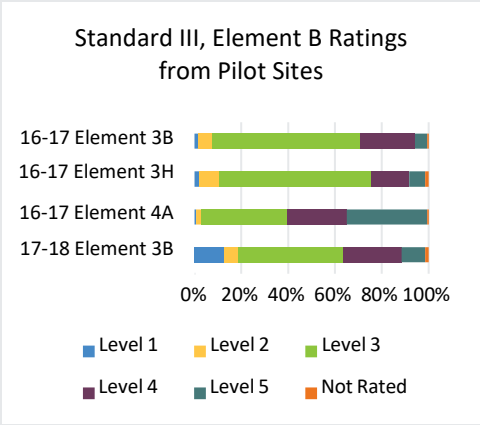


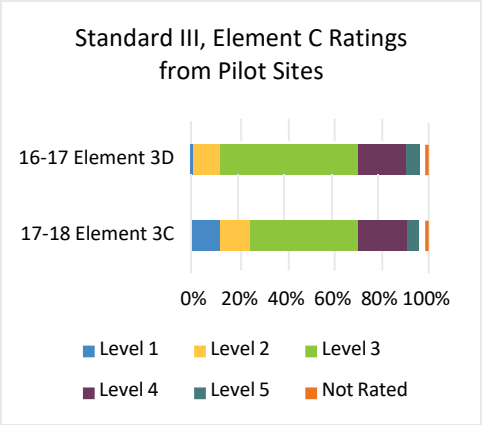


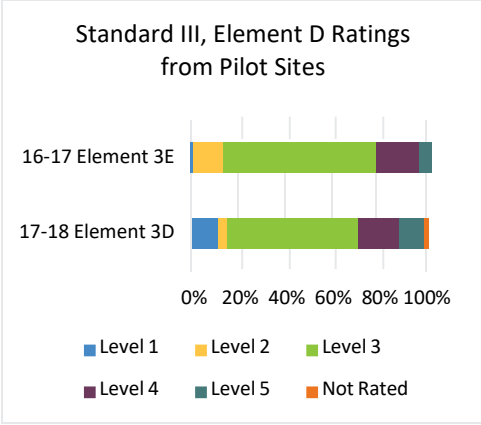


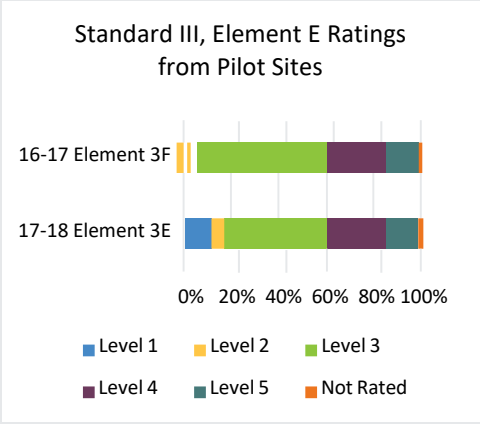

Standard II, Element D Ratings from Pilot Sites
16-17 Element 2E
17-18 Element 2D
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated

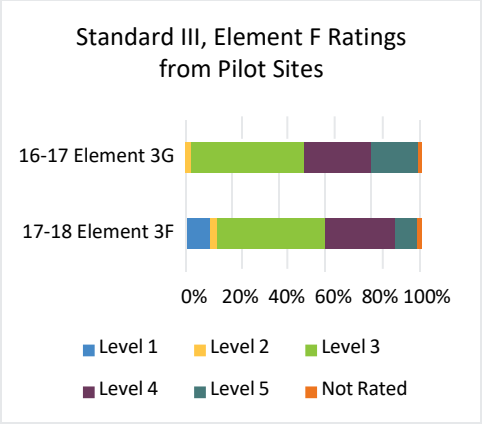

Standard III, Element A Ratings From Pilot Sites
16-17 Element 3C
16-17 Element 3A
17-18 Element 3A
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated

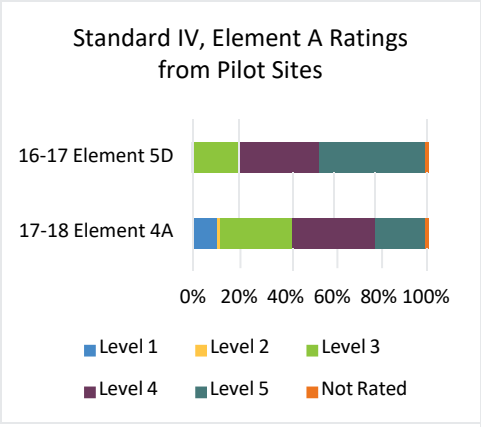

Standard III, Element B Ratings from Pilot Sites
16-17 Element 3B
16-17 Element 3H
16-17 Element 4A
17-18 Element 3B
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated

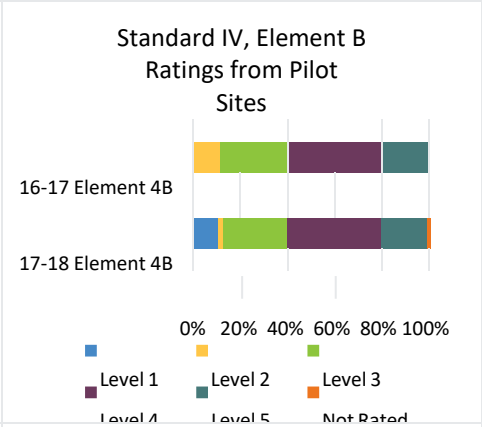

Standard III, Element C Ratings from Pilot Sites
16-17 Element 3D
17-18 Element 3C
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated

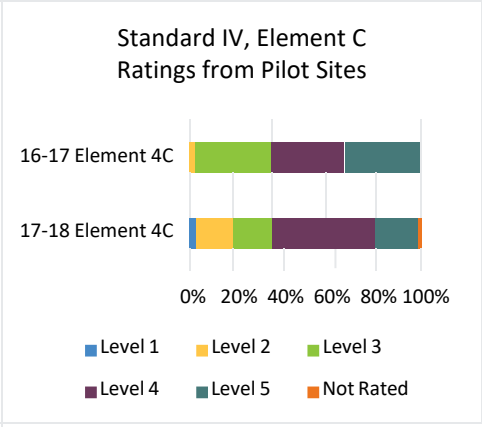

Standard III, Element D Ratings from Pilot Sites
16-17 Element 3E
17-18 Element 3D
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated

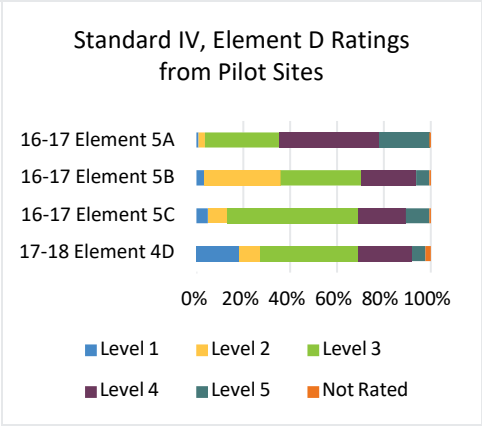

Standard III, Element E Ratings from Pilot Sites
16-17 Element 3F
17-18 Element 3E
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated


Standard III, Element F Ratings from Pilot Sites
16-17 Element 3G
17-18 Element 3F
0% 20% 40% 60% 80% 100%
Level 1
Level 2
Level 3
Level 4
Level 5
Not Rated







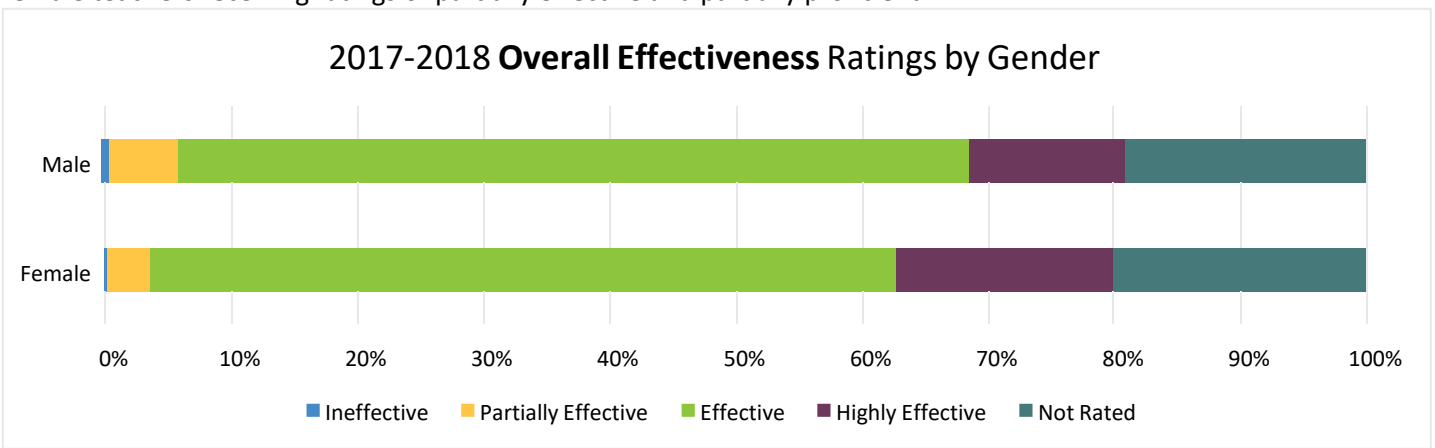


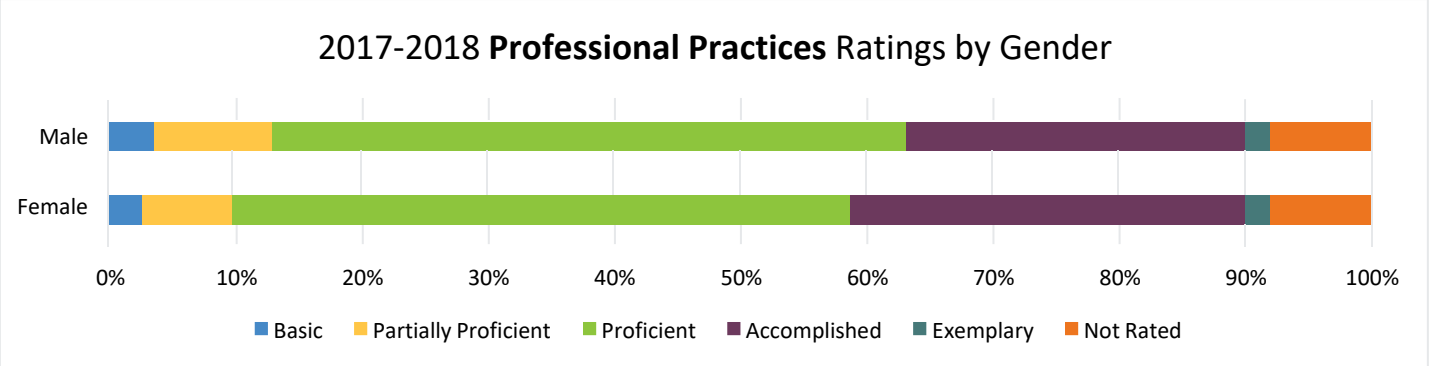
DISTRIBUTIONS OF SCORES BY TEACHER DEMOGRAPHIC GROUP

This section examines differences in the pilot ratings that may exist among teacher demographic groups of gender and race/ethnicity.

Gender

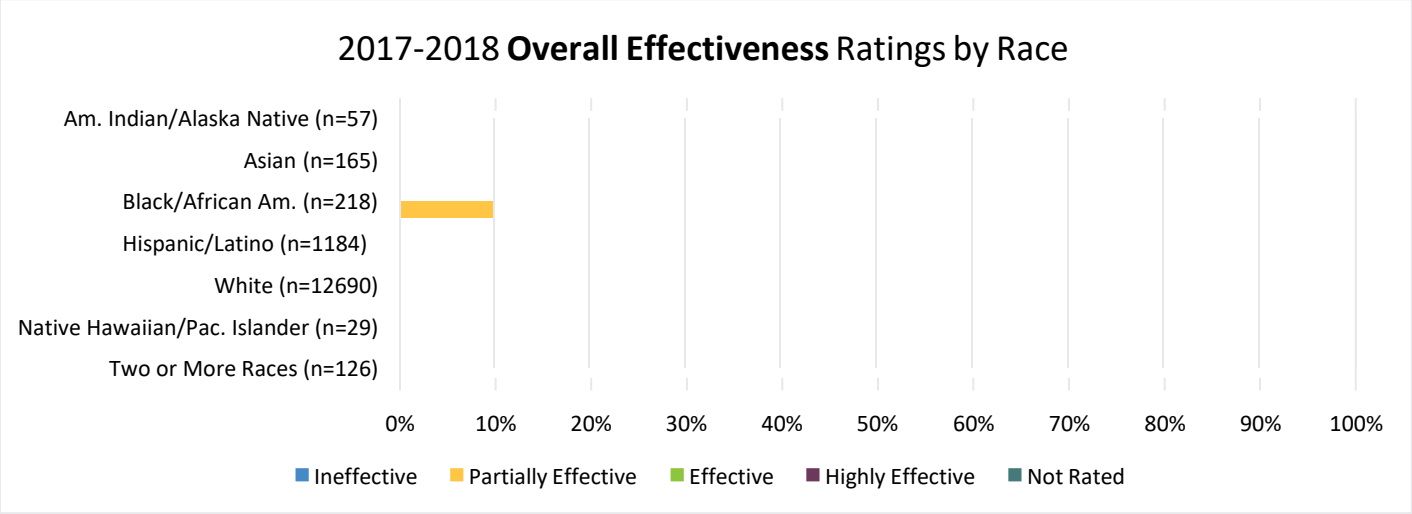
The gender patterns for overall effectiveness ratings and professional practices ratings are similar, with slightly larger proportions of female teachers receiving ratings of highly effective and accomplished and slightly smaller proportions of female teachers receiving ratings of partially effective and partially proficient.

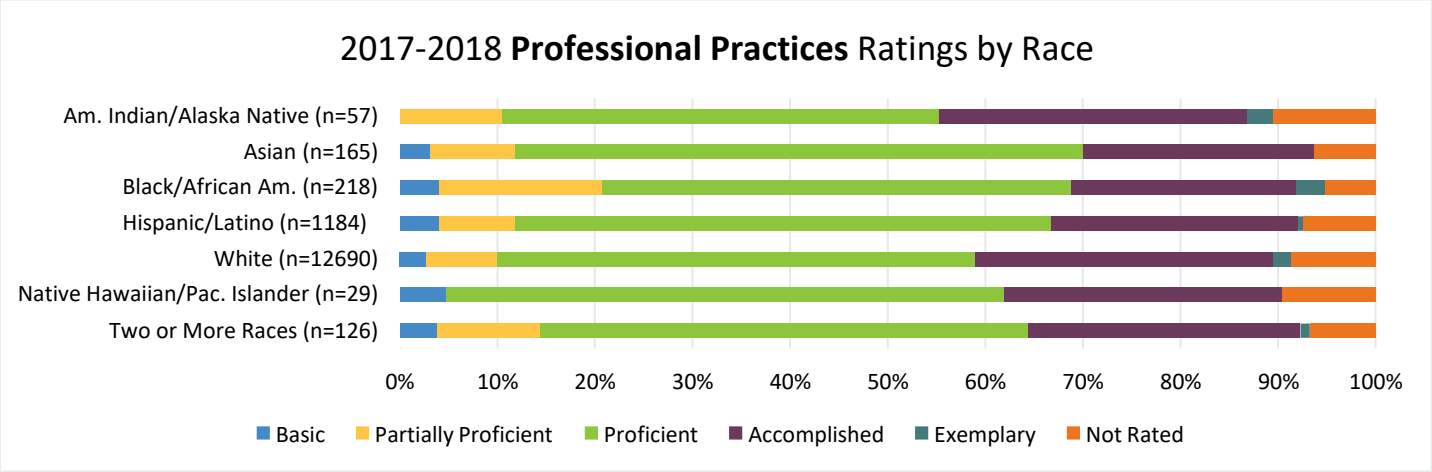




Race/Ethnicity

The following graphs comparing ratings by federal race/ethnicity categories include information about group size using "n=" to indicate the number of teachers in a specific group. This information is included because the sizes of these groups varies substantially and, in some cases, is quite small. Results should be interpreted cautiously for small groups, such as Native Hawaiian/Pacific Islander, in the following charts because of the disproportionate impact of one individual in a small group compared to a larger group.



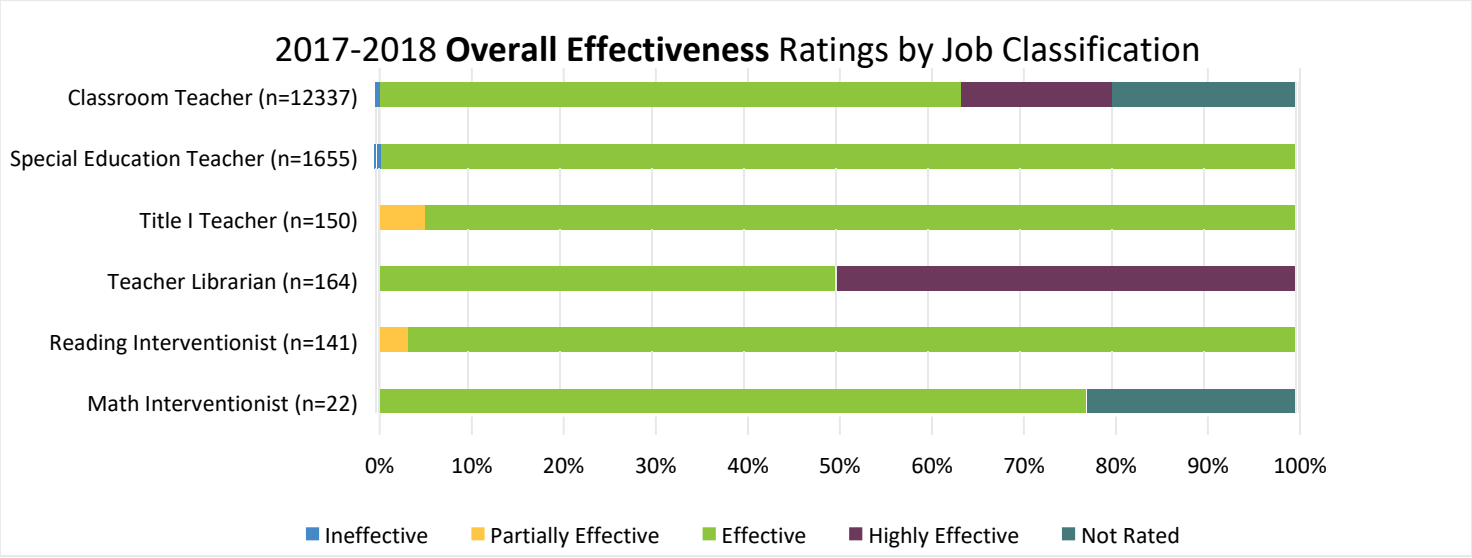


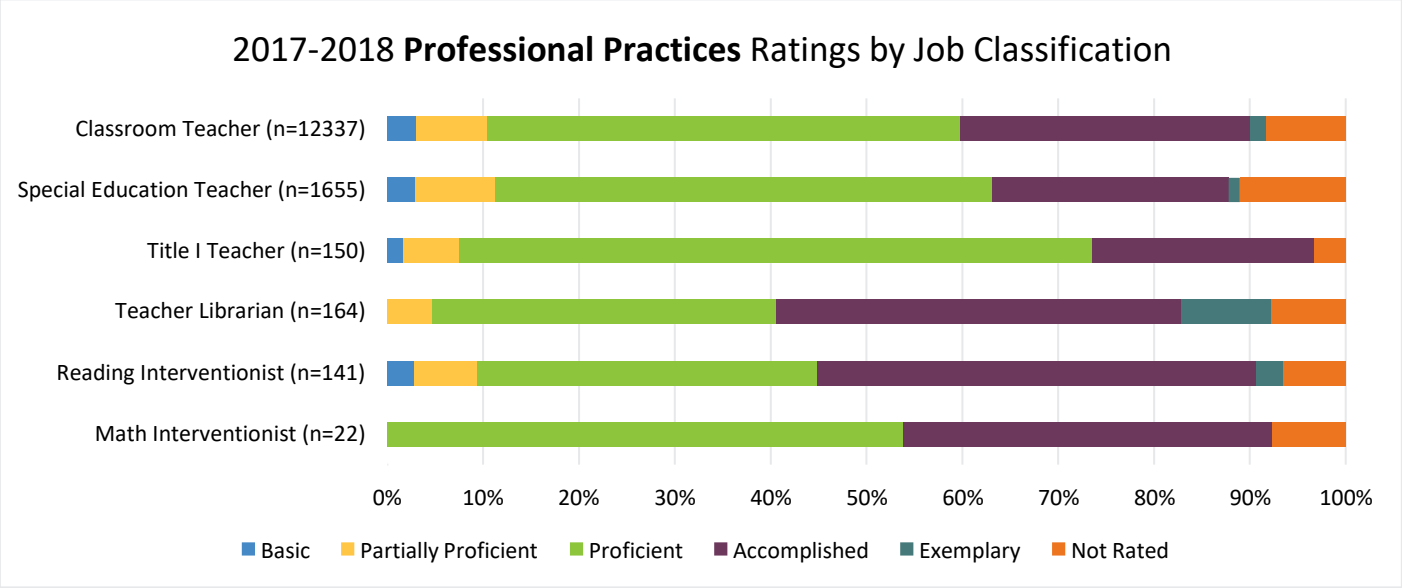
DISTRIBUTIONS OF SCORES BY TEACHER EMPLOYMENT CATEGORIES

This section presents overall ratings for teachers who fall into different employment groups, as reported by districts in the annual staff assignment human resources collection. The specific employment groups considered in this section include teachers' specific job classification, content area taught, grade level in which they teach, length of time they have taught (i.e., 3 or fewer years of experience and 4 or more years of experience), and their status as non-probationary or probationary. In most cases, there are differences in the ratings distributions for the employment categories represented in the chart. One area that appears to be consistent is grade level, with similar rating distributions for elementary, middle and high school teachers. Experienced teachers and non-probationary teachers outperform novice teachers and probationary teachers.

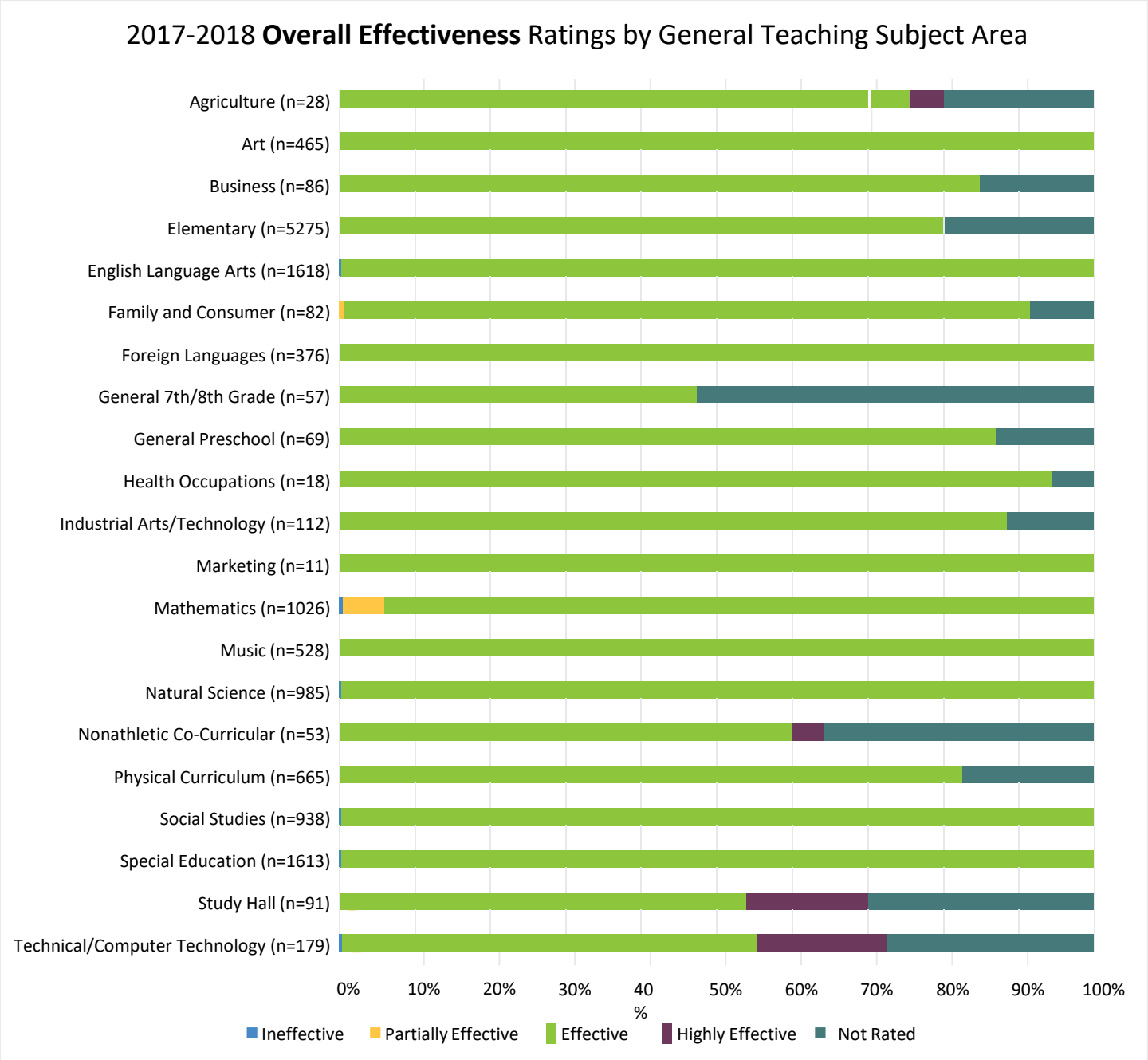
The following graphs for job classification and general teaching subject area include information about group size using "n=" to indicate the number of teachers in a specific group. This information is included because the sizes of these groups varies substantially and, in some cases, is quite small. Results should be interpreted cautiously for small groups, such as Math Interventionist, Agriculture, Health Occupations, and Marketing, in the following chart, as the ratings distribution may be quite different from that of a larger group had one been available.

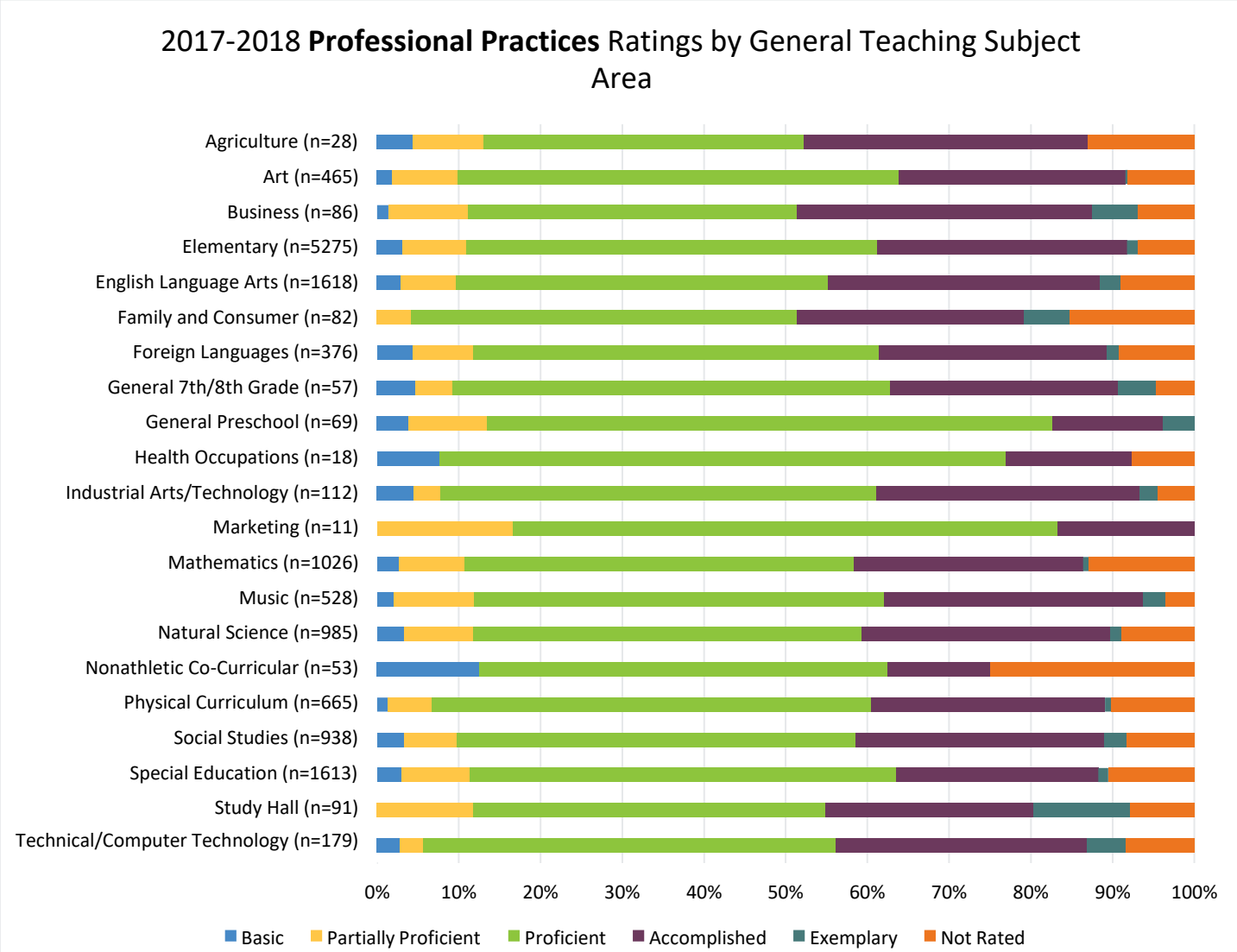
Job Classification





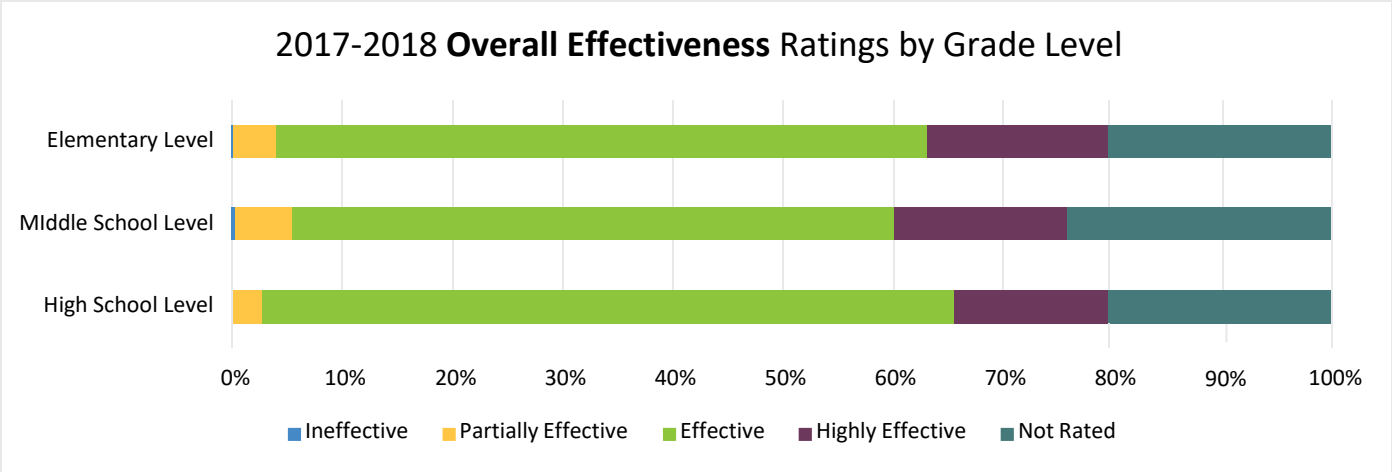
Content Area Taught

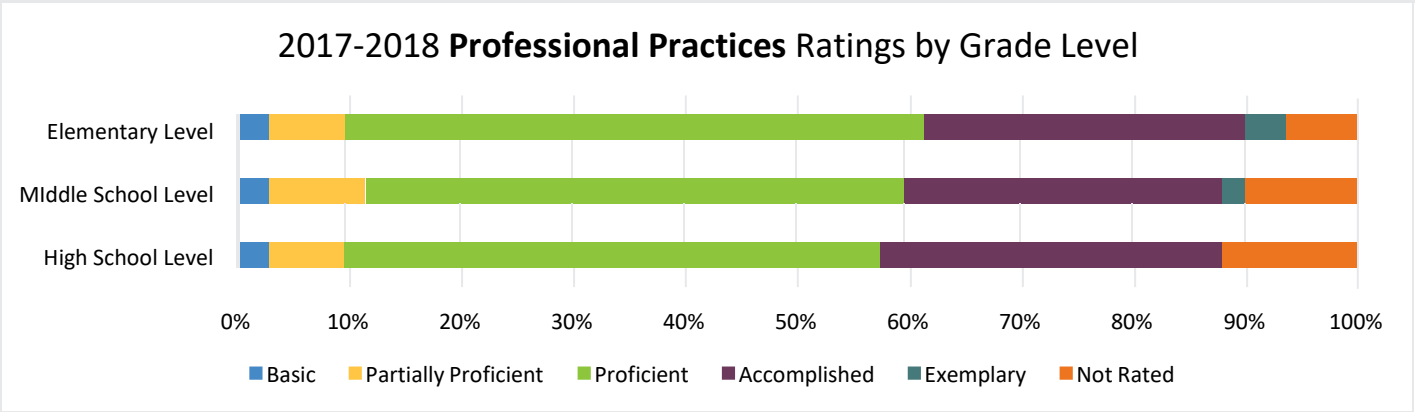




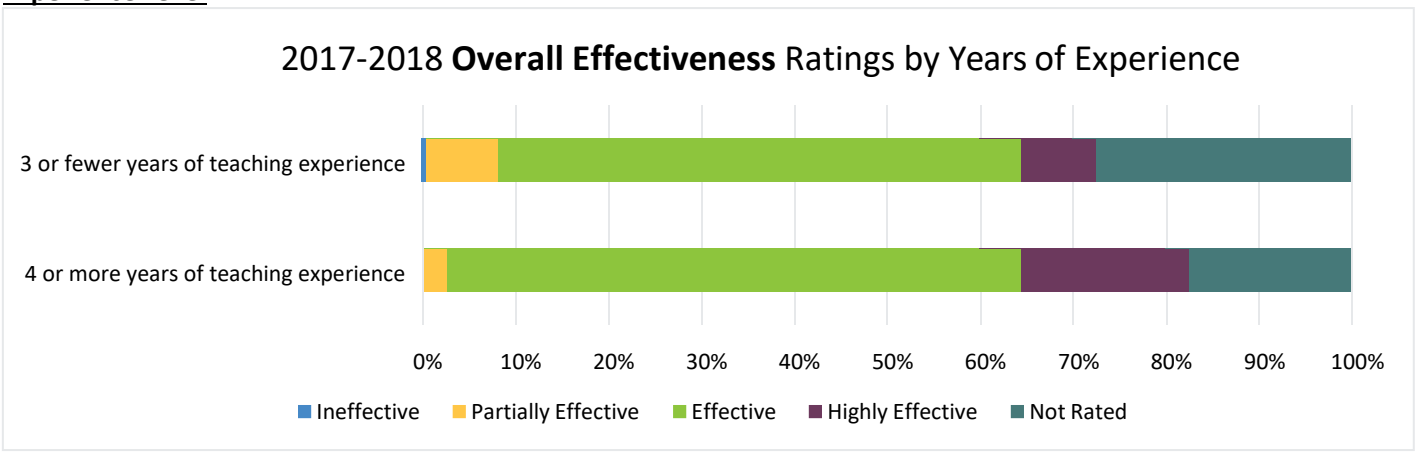
Grade Level Taught

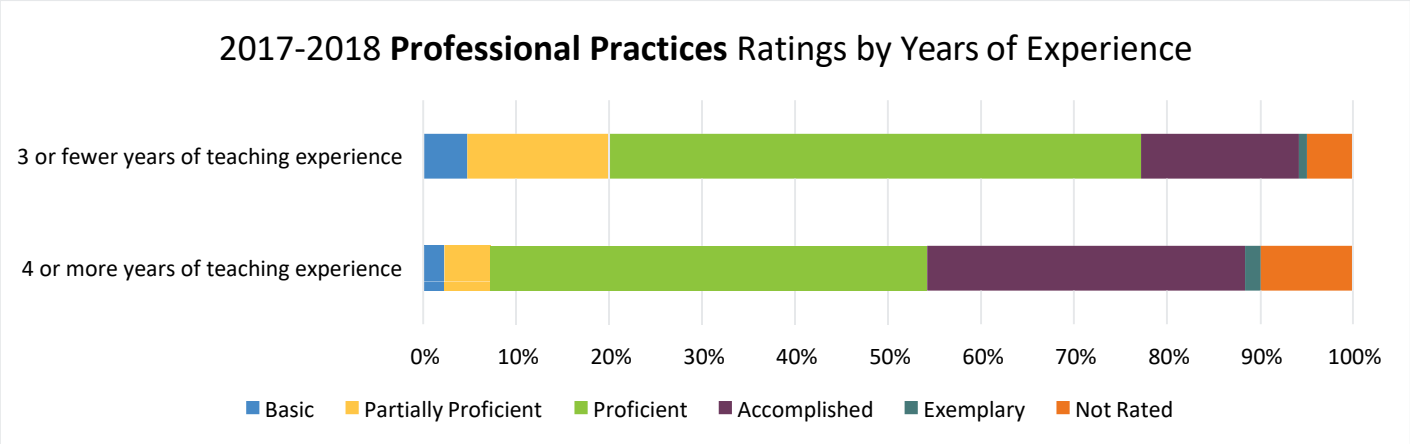
The following graphs display comparisons of ratings by grade level groupings in which teachers taught. In these graphs, the elementary level is comprised of pre-kindergarten through grade 5, the middle school level is comprised of grades 6 through 8, and the high school level is comprised of grades 9 through 12. In the graphs, a teacher is included in each grade level grouping in which they taught.



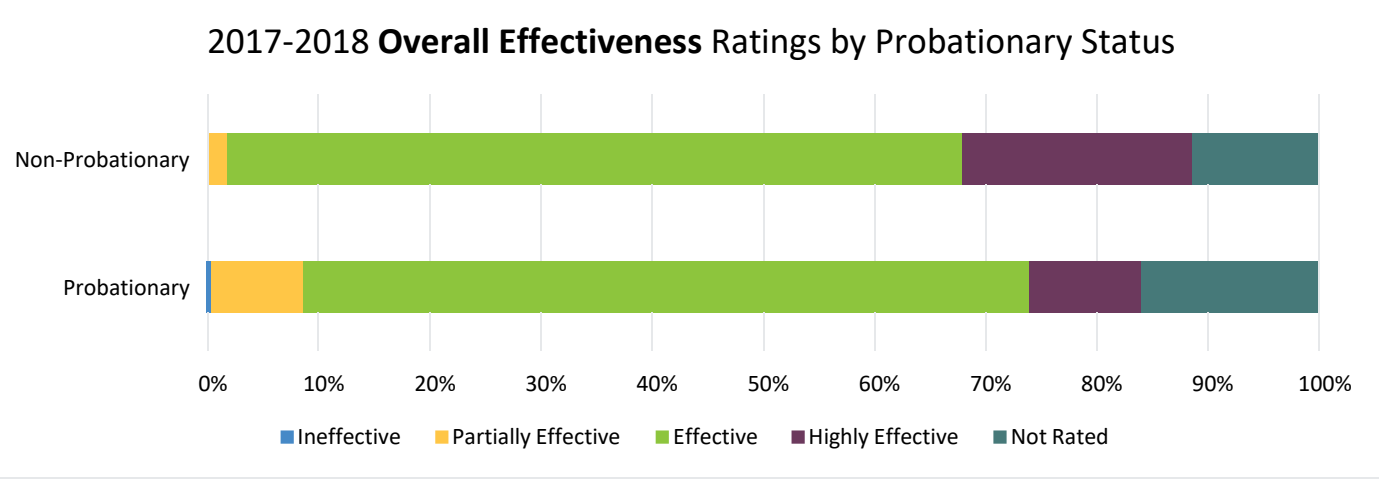


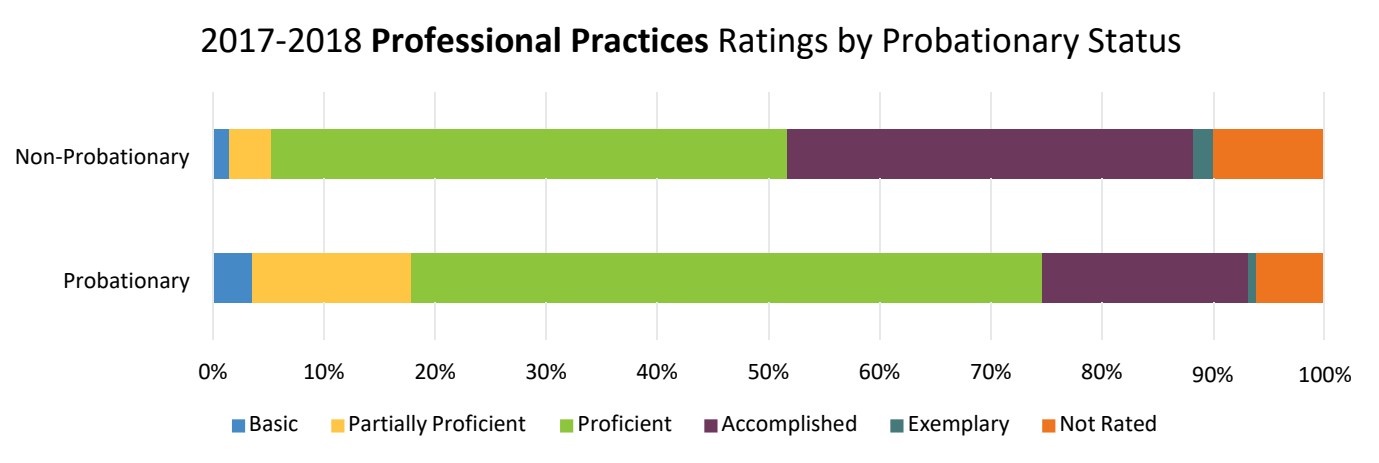
Experience Level





Tenure Status





This concludes the presentation of teacher evaluation ratings. The revised SMES for teachers resulted in more variation of overall effectiveness ratings and professional practices ratings with fewer teachers clustered at the highest rating levels. This shift reflects the intentional increased rigor of the rubric and scoring process that stakeholders requested.

EDUCATOR SURVEY RESPONSES

In this section of the pilot report, we present the results of the perception surveys. Surveys were administered online to teachers, principals/evaluators, and district administrators. For those using the Colorado Performance Management System (RANDA), the feedback surveys were sent directly to educators' district emails on file in that system by role. For those employed in districts not utilizing the system, the survey link was sent to district leadership to disseminate to their staff in appropriate roles. The teacher respondent group consisted of 5341 teachers for the pre-survey and 4279 teachers for the post-survey; the principal/evaluator respondent group consisted of 826 on the pre-survey and 379 on the post-survey; and the administrator group consisted of 62 on the pre-survey and 48 on the post-survey.

The pre-pilot survey, administered in September 2017, asked respondents to reflect on their experiences with the 2016-2017 SMES for teachers, i.e., the former evaluation process and rubric used the previous school year. The post-pilot survey, administered in April 2018, asked respondents to reflect on their experiences with the revised SMES for teachers used during the pilot in 2017-2018.

The survey was a comprehensive instrument that solicited feedback on several dimensions of the SMES for teachers. It contained items specific to the teacher rubric, Measures of Student Learning (MSL), scoring, local implementation, and the overall evaluation system, as well as respondent demographic information. The same survey was administered to each respondent group, with minor wording changes to reflect the respondent role in the evaluation process.

Responses to items administered on a 4-point Likert rating scale (e.g., Not at all satisfied, Slightly Satisfied, Satisfied, Extremely Satisfied) or 5-point Likert rating scale (e.g., Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) were consolidated for purposes of reporting here. The tables presented here contain the percentage of favorable responses, which, for both the 4-point and 5-point scales, includes a response in either of the last two categories (Satisfied and Extremely Satisfied, or Agree and Strongly Agree, in these examples).

Responses to other items are reported as means where the range for each item is between 1 and 5. Items reported with means have response options that are not on a clear Likert scale (for example, less than 1 hour, 1-2 hours, 2.1-3 hours, 3.1-4 hours, more than 4 hours). For these other items (see tables 2, 4, and 6), it is important to consider not only the change from pre- to post-test, but also the direction of the change. For example, the item regarding time burden on teachers is rated on the preceding scale and, as such, a decrease in the amount of time is more desirable. For items where the change is statistically or practically significant, the direction of the change is labeled ("Increase" or "Decrease") and highlighted green to signal when the change is in the desirable direction.

Results are analyzed separately for teachers, principal/evaluators, and administrators. Item-level reports for each item with both pre- and post-test results are included in the pages that follow. Reported for each item is the pre-test mean, post-test mean, percentage point change, statistical significance based on a *t*-test, and effect size (or practical significance) based on Cohen's *d*. Statistical significance and effect size interpretation appear beneath each table. To better understand the impact of the revised SMES for teachers, survey items most associated with the foci of the revisions are discussed. (For more details, please see tables of all survey results presented below.)

There were several positive responses to the pilot shared by teachers, principal/evaluators, and administrators on items pertaining to the length and clarity of the rubric, more rigorous scoring methods, and overall support for the SMES for teachers. A few specific survey item responses that changed in a desirable direction with statistical and practical significance are highlighted below:

- The reduction in the length of the rubric, the number of elements and the number of discrete professional practices reflected in the rubric were very well-received by teachers (favorable responses increased by 32.6,

39.2, and 39.5 percentage points, respectively); principal/evaluators (favorable responses increased by 52.9, 53.1, and 59.5 percentage points, respectively); and administrators (favorable responses increased by 63.9, 69.2, and 69.2 percentage points, respectively).

- Similarly, all educator groups reported statistically and practically significant increases in their satisfaction with the improved clarity of language, reduction in the extent of redundancy, the understanding of the meaning of each professional practice, and their ability to use the rubric as a reflection tool. Teacher responses to these items increased 18.7, 31.6, 14.9, and 13.4 percentage points; principal/evaluator responses to these items increased 32.8, 52.9, 20.0, and 24.6 percentage points; and administrator responses to these items increased 48.5, 61.0, 50.2, and 32.1 percentage points, respectively.
- A statistically and practically significant reduction in the time burden for teachers (from an average of 3.5 hours to an average of 3.2 hours) and principal/evaluators (from an average of 3.10 hours to an average of 2.67 hours) was reported. For administrators, the average time was reduced from an average of 3.24 hours to an average of 2.85 hours, reflecting a small effect size but no statistical significance.
- Across the board, there also were statistically significant (and practically significant for principal/evaluators and administrators) increases in satisfaction with the rigor for earning overall ratings of Effective and Highly Effective. In each case, nearly two-thirds or more of respondents viewed the overall rating scoring changes favorably, with teachers' satisfaction increasing to 68.8% for Effective and 71.5% for Highly Effective, principal/evaluators' satisfaction increasing to 65.7% for Effective and 75.8% for Highly Effective, and administrators' satisfaction increasing to 71.1% for Effective and 70.5% for Highly Effective.
- Perceptions of overall support for the rubric, the teacher evaluation system, and the Measures of Student Learning (MSL) portion of the system have increased significantly for all educators, although the increase and the favorable percentages are much higher for principal/evaluators and administrators than they are for teachers. Additionally, principal/evaluators and administrators reported significantly more favorable perceptions about the overall evaluation rating accurately reflecting a teacher's practice.

TEACHER SURVEY RESPONSES

Table 1: Teacher responses to Likert-type items

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: TEACHER RESPONSES TO SURVEY	Pre-Test % Favorable Response (N=5341)	Post-Test % Favorable Response (N=4279)	Percentage Point Change From Pre to Post	Statistically Significant Change	Effect Size (Cohen's d)
Provided me with clear expectations for my teaching	54.8%	57.8%	3.0	*	Small
Encompassed all aspects of quality teaching	48.5%	49.7%	1.2		
Identified areas of strength	70.0%	67.0%	-3.1	***	
Identified areas that need improvement	65.2%	63.0%	-2.2	**	
Provided an accurate assessment of my performance	41.7%	39.7%	-1.9	**	
Focused on the aspects of my work that will affect student learning	48.8%	51.3%	2.5		
Resulted in improving student growth	36.0%	39.2%	3.2		
Provided actionable feedback to the person being evaluated	46.8%	51.4%	4.5	**	
Provided timely feedback to the person being evaluated	49.3%	55.3%	6.0	***	
Was useful to me in making instructional decisions	36.6%	40.3%	3.7		
Influenced my practice as a teacher	45.6%	47.0%	1.4		
Helped improve my instructional practice	40.1%	42.3%	2.2		
Set high standards for the person being evaluated	63.3%	65.5%	2.2		
Set a high bar for earning an overall rating of Effective	66.3%	68.8%	2.5	**	
Set a high bar for earning an overall rating of Highly Effective	69.6%	71.5%	1.9	*	
Provided a system where evaluators provided support to educators	41.9%	44.4%	2.5		
The length of the rubric	27.7%	60.3%	32.6	***	Medium
The number of Teacher Quality Standards reflected in the rubric (i.e., reduced from 5 standards to 4)	42.7%	68.9%	26.2	***	Medium
The number of elements reflected in the rubric (i.e., reduced from 27 elements to 17)	29.7%	68.9%	39.2	***	Large
The number of discrete professional practices reflected in the rubric (i.e., reduced from approximately 300 practices to 167)	28.1%	67.6%	39.5	***	Large
The performance category labels (i.e., previously Basic, Partially Proficient, etc. and now is Level I - Level V)	60.6%	67.4%	6.8	***	Small
The reduction of professional practices for elementary and secondary language arts	47.5%	73.4%	25.9	***	Small
The reduction of professional practices for mathematics educators	49.9%	72.9%	23.0	***	Small
The progression of rigor across performance categories	53.7%	67.5%	13.8	***	Small
The level of language clarity	43.9%	62.7%	18.7	***	Small
The extent of redundancies	33.7%	65.3%	31.6	***	Medium
Your understanding of the meaning of each professional practice	49.7%	64.6%	14.9	***	Small
Your ability to use the rubric as a reflection tool	44.6%	58.0%	13.4	***	Small
The shared understanding with your evaluator(s)	58.5%	66.4%	7.9	***	
The targets established for my Measures of Student Learning (MSLs) were rigorous	66.0%	68.5%	2.4	**	

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: TEACHER RESPONSES TO SURVEY	Pre-Test % Favorable Response (N=5341)	Post-Test % Favorable Response (N=4279)	Percentage Point Change From Pre to Post	Statistically Significant Change	Effect Size (Cohen's d)
The targets established for my MSLs were attainable	60.5%	61.7%	1.2		
I used the MSL data from my evaluation to inform my instruction	55.3%	56.3%	1.0		
There was a high bar to earn a rating of Expected	69.4%	72.1%	2.7	**	
There was a high bar to earn a rating of More than Expected	72.6%	75.0%	2.4	**	
My principal/evaluator(s) used data from my evaluation to identify professional supports for me	51.1%	50.2%	-0.9		
The rubric helped me identify relevant professional development to address my specific areas for improvement	44.8%	48.5%	3.7	*	
The rubric helped me identify specific professional goals to improve my practice	52.2%	55.1%	2.9		
My principal/evaluator(s) used summative evaluation data to inform targeted PD opportunities for my building	38.4%	41.9%	3.5	*	
The evaluation process provided opportunities to have meaningful conversations about my practice with my principal/evaluator(s)	56.4%	57.1%	0.7		
I felt supportive of the teacher evaluation rubric	34.9%	41.5%	6.6	***	
I felt supportive of the MSL portion of the evaluation system	32.9%	37.9%	5.0	***	
I felt supportive of the teacher evaluation system overall	34.0%	39.8%	5.8	***	
In your opinion, my overall rating is an accurate reflection of my practice as a teacher.	70.2%	66.5%	-3.7	***	

*Statistical Significance legend: *p<.05, **p<.01, ***p<.001*

Cohen's d legend: no effect<.2, .2<=small effect<.5, .5<=medium effect<.8, large effect>=.8

Table 2: Teacher responses to other items

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: TEACHER RESPONSES TO SURVEY	Pre-Test Mean (N=5341)	Post-Test (N=4279)	Statistically Significant Change	Effect Size (Cohen's d)	Direction
Time burden on teachers	3.5	3.2	***	Small	Decrease
Time burden on principals/assistant principals	3.9	3.6	***	Small	Decrease
Total Time on Self-assessment	2.3	2.2	***		Decrease
Total Time on Goal-setting	2.4	2.2	***		Decrease
Total Time on Post-observation conversations	2.0	1.9	***		Decrease
Total Time on Mid-year review	1.8	1.7	***		Decrease
Total Time on End-of-year review	1.9	1.7	***		Decrease
Was the time spent on implementation time well spent?	2.5	2.7	***		Increase
The scoring approach for professional practice	2.7	2.9	***		Increase
The scoring approach for MSL	2.7	2.8	***		Increase
The scoring approach for the overall effectiveness rating	2.7	2.8	***		Increase
In your opinion, how effective was the State Model Evaluation System for Teachers during this current school year?	2.07	2.2	***	Small	Increase

*Statistical Significance legend: *p<.05, **p<.01, ***p<.001*

Cohen's d legend: no effect<.2, .2<=small effect<.5, .5<=medium effect<.8, large effect>=.8

PRINCIPAL/EVALUATOR SURVEY RESPONSES

Table 3: Principal/Evaluator responses to Likert-type items

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: PRINCIPAL/EVALUATOR RESPONSES TO SURVEY	Pre-Test % Favorable Response (N=826)	Post-Test % Favorable Response (N=379)	Percentage Point Change From Pre to Post	Statistically Significant Change	Effect Size (Cohen's d)
Provided clear expectations for teachers	55.5%	71.7%	16.1	***	Small
Encompassed all aspects of quality teaching	54.9%	60.3%	5.4	*	
Identified areas of strength	74.9%	81.7%	6.8	**	Small
Identified areas that need improvement	68.7%	76.3%	7.6	**	
Provided an accurate assessment of teacher performance	31.7%	47.5%	15.8	***	Small
Focused on the aspects of a teacher's work that affected student learning	51.4%	68.6%	17.2	***	Small
Resulted in improving student growth	27.1%	41.8%	14.6	***	Small
Provided actionable feedback to the person being evaluated	48.8%	65.0%	16.2	***	Small
Provided timely feedback to the person being evaluated	47.6%	64.3%	16.8	***	Small
Was useful to me in helping a teacher make instructional decisions	38.7%	52.1%	13.4	***	Small
Influenced teacher practice in my district	34.9%	47.9%	13.0	***	Small
Helped improve instructional practices in my district	30.7%	43.4%	12.8	***	Small
Set high standards for the person being evaluated	61.8%	74.2%	12.4	***	Small
Set a high bar for earning an overall rating of Effective	50.8%	65.7%	15.0	***	Small
Set a high bar for earning an overall rating of Highly Effective	58.4%	75.8%	17.5	***	Medium
Provided a system where evaluators provided support to educators	49.3%	62.4%	13.1	***	Small
The length of the rubric	22.7%	75.6%	52.9	***	Large
The number of Teacher Quality Standards reflected in the rubric (i.e. reduction from 5 standards to 4)	42.5%	83.6%	41.1	***	Large
The number of elements reflected in the rubric (i.e., reduction from 27 elements to 17)	25.7%	78.9%	53.1	***	Large
The number of discrete professional practices reflected in the rubric (i.e. reduction from approximately 300 practices to 167)	23.6%	83.1%	59.5	***	Large
The performance category labels (i.e., previously Basic, Partially Proficient, etc. and is now Level 1 - Level 5)	64.3%	83.7%	19.4	***	Medium
The additional/reduction of professional practices for elementary and secondary language arts	48.1%	87.3%	39.3	***	Large
The additional/reduction of professional practices for mathematics educators	49.0%	86.8%	37.8	***	Large
The progression of rigor across performance categories	54.3%	80.1%	25.8	***	Medium
The level of language clarity	37.3%	70.1%	32.8	***	Medium
The extent of redundancies	26.7%	79.7%	52.9	***	Large
Your understanding of the meaning of each professional practice	49.0%	69.1%	20.0	***	Small
Your ability to use the rubric as a reflection tool	50.6%	75.2%	24.6	***	Medium
The shared understanding with your teachers	48.4%	73.3%	24.8	***	Medium
The targets established for Measures of Student Learning (MSLs) were rigorous	43.5%	53.6%	10.1	***	Small

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: PRINCIPAL/EVALUATOR RESPONSES TO SURVEY	Pre-Test % Favorable Response (N=826)	Post-Test % Favorable Response (N=379)	Percentage Point Change From Pre to Post	Statistically Significant Change	Effect Size (Cohen's d)
The targets established for MSLs were attainable	55.3%	71.2%	15.9	***	Small
Teachers used the MSL data from their evaluation to inform instruction	34.5%	50.1%	15.6	***	Small
There was a high bar to earn a rating of Expected	43.7%	57.3%	13.6	***	Small
There was a high bar to earn a rating of More than Expected	51.7%	64.7%	12.9	***	Small
I used data from evaluations to identify professional supports for my staff	43.6%	64.1%	20.4	***	Small
The rubric helped me identify relevant professional development to address specific areas for improvement in my staff	34.7%	53.2%	18.4	***	Small
The rubric helped me identify specific professional goals for my staff	39.6%	56.9%	17.3	***	Small
I used summative evaluation data to inform targeted PD opportunities for my building	27.5%	45.1%	17.6	***	Small
The evaluation process provided opportunities to have meaningful conversations about professional practice with my staff	50.5%	69.9%	19.4	***	Small
I felt supportive of the teacher evaluation rubric	42.5%	69.0%	26.6	***	Small
I felt supportive of the MSL portion of the evaluation system	36.3%	55.7%	19.4	***	Small
I felt supportive of the teacher evaluation system overall	41.3%	64.9%	23.6	***	Small
In your opinion, did the overall ratings of your educators give an accurate reflection of teacher practice?	65.9%	73.5%	7.6	*	Small

*Statistical Significance legend: *p<.05, **p<.01, ***p<.001*

Cohen's d legend: no effect<.2, .2<=small effect<.5, .5<=medium effect<.8, large effect>=.8

Table 4: Principal/Evaluator responses to other items

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: PRINCIPAL/EVALUATOR RESPONSES TO SURVEY	Pre-Test Mean (N=826)	Post-Test (N=379)	Statistically Significant Change	Effect Size (Cohen's d)	Direction
Time burden on teachers	3.10	2.67	***	Small	Decrease
Time burden on principals/assistant principals	4.05	3.76	***	Small	Decrease
Total Time on Self-assessment	1.89	1.60	***	Small	Decrease
Total Time on Goal-setting	2.20	1.95	**	Small	Decrease
Total Time on Post-observation conversations	2.45	2.47			No change
Total Time on Mid-year review	2.33	2.31			No change
Total Time on End-of-year review	2.43	2.38			No change
Was the time spent on implementation time well spent?	2.81	3.26	***	Small	Increase
The scoring approach for professional practice	2.64	3.33	***	Medium	Increase
The scoring approach for MSL	2.51	3.03	***	Small	Increase
The scoring approach for the overall effectiveness rating	2.52	3.26	***	Medium	Increase
In your opinion, how effective was the State Model Evaluation System for Teachers during this current school year?	2.33	2.47	**	Small	Increase

*Statistical Significance legend: *p<.05, **p<.01, ***p<.001*

Cohen's d legend: no effect<.2, .2<=small effect<.5, .5<=medium effect<.8, large effect>=.8

DISTRICT ADMINISTRATOR SURVEY RESPONSES

Table 5: District administrator responses to Likert-type items

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: ADMINISTRATOR RESPONSES TO SURVEY	Pre-Test % Favorable Response (N=62)	Post-Test % Favorable Response (N=48)	Percentage Point Change From Pre to Post	Statistically Significant Change	Effect Size (Cohen's d)
Provided clear expectations for teachers	69.8%	80.4%	10.6		Medium
Encompassed all aspects of quality teaching	56.6%	63.0%	6.4		
Identified areas of strength	73.1%	82.6%	9.5		Small
Identified areas that need improvement	66.0%	73.9%	7.9		Small
Provided an accurate assessment of teacher performance	22.6%	52.2%	29.5	***	Medium
Focused on the aspects of a teacher's work that affected student learning	57.7%	73.9%	16.2		Small
Resulted in improving student growth	22.6%	43.5%	20.8	*	Small
Provided actionable feedback to the person being evaluated	54.7%	78.3%	23.5		Small
Provided timely feedback to the person being evaluated	45.3%	65.2%	19.9	*	Small
Was useful to me in helping a teacher make instructional decisions	26.4%	54.3%	27.9	**	Medium
Influenced teacher practice in my district	39.6%	52.2%	12.6		Small
Helped improve instructional practices in my district	22.6%	47.8%	25.2	**	Medium
Set high standards for the person being evaluated	67.3%	75.6%	8.2		Small
Set a high bar for earning an overall rating of Effective	42.3%	71.1%	28.8	**	Medium
Set a high bar for earning an overall rating of Highly Effective	41.2%	70.5%	29.3	**	Medium
Provided a system where evaluators provided support to educators	55.8%	77.8%	22.0		Small
The length of the rubric	12.2%	76.2%	63.9	***	Large
The number of Teacher Quality Standards reflected in the rubric (i.e. reduction from 5 standards to 4)	27.1%	76.7%	49.7	***	Large
The number of elements reflected in the rubric (i.e., reduction from 27 elements to 17)	12.2%	81.4%	69.2	***	Large
The number of discrete professional practices reflected in the rubric (i.e. reduction from approximately 300 practices to 167)	12.2%	81.4%	69.2	***	Large
The performance category labels (i.e., previously Basic, Partially Proficient, etc. and is now Level I - Level V)	69.4%	81.4%	12.0		Small
The additional/reduction of professional practices for elementary and secondary language arts	34.1%	87.5%	53.4	***	Large
The additional/reduction of professional practices for mathematics educators	33.3%	86.8%	53.5	***	Large
The progression of rigor across performance categories	50.0%	78.6%	28.6	**	Medium
The level of language clarity	30.6%	79.1%	48.5	***	Large
The extent of redundancies	20.4%	81.4%	61.0	***	Large
Your understanding of the meaning of each professional practice	26.5%	76.7%	50.2	***	Large
Your ability to use the rubric as a reflection tool	46.9%	79.1%	32.1	***	Medium
The shared understanding with your teachers	55.1%	81.0%	25.9	**	Medium
The targets established for Measures of Student Learning (MSLs) were rigorous	37.3%	52.3%	15.0	*	Medium

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: ADMINISTRATOR RESPONSES TO SURVEY	Pre-Test % Favorable Response (N=62)	Post-Test % Favorable Response (N=48)	Percentage Point Change From Pre to Post	Statistically Significant Change	Effect Size (Cohen's d)
The targets established for MSLs were attainable	64.7%	68.2%	3.5		
Teachers used the MSL data from their evaluation to inform instruction	28.0%	50.0%	22.0	*	Medium
There was a high bar to earn a rating of Expected	29.4%	61.9%	32.5	*	Medium
There was a high bar to earn a rating of More than Expected	37.3%	69.8%	32.5	**	Medium
I used data from evaluations to identify professional supports for my staff	52.9%	52.3%	-0.7		Small
The rubric helped me identify relevant professional development to address specific areas for improvement in my staff	45.1%	52.3%	7.2		Small
The rubric helped me identify specific professional goals for my staff	64.0%	70.5%	6.5		Small
I used summative evaluation data to inform targeted PD opportunities for my building	39.2%	40.5%	1.3		Small
The evaluation process provided opportunities to have meaningful conversations about professional practice with my staff	74.5%	79.1%	4.6		Small
I felt supportive of the teacher evaluation rubric	50.0%	68.3%	18.3		Small
I felt supportive of the MSL portion of the evaluation system	35.4%	56.1%	20.7		Small
I felt supportive of the teacher evaluation system overall	52.1%	65.9%	13.8		Small
In your opinion, did the overall ratings of your educators give an accurate reflection of teacher practice?	47.9%	68.3%	20.4	*	Small

*Statistical Significance legend: *p<.05, **p<.01, ***p<.001*

Cohen's d legend: no effect<.2, .2<=small effect<.5, .5<=medium effect<.8, large effect>=.8

Table 6: District administrator responses to other items

REVISED TEACHER EVALUATION RUBRIC PILOT TEST: ADMINISTRATOR RESPONSES TO SURVEY	Pre-Test Mean (N=62)	Post-Test (N=48)	Statistically Significant Change	Effect Size (Cohen's d)	Direction
Time burden on teachers	3.24	2.85		Small	Decrease
Time burden on principals/assistant principals	4.47	3.71	***	Medium	Decrease
Total Time on Self-assessment	1.82	1.76			No change
Total Time on Goal-setting	2.02	2.00			No change
Total Time on Post-observation conversations	2.42	2.21			No change
Total Time on Mid-year review	2.30	2.00		Small	Decrease
Total Time on End-of-year review	2.53	2.05		Small	Decrease
Was the time spent on implementation time well spent?	3.08	3.71	**	Medium	Increase
The scoring approach for professional practice	2.88	3.40	**	Medium	Increase
The scoring approach for MSL	2.78	3.12		Small	Increase
The scoring approach for the overall effectiveness rating	2.78	3.29	*	Small	Increase
In your opinion, how effective was the State Model Evaluation System for Teachers during this current school year?	2.19	2.63	**	Medium	Increase

*Statistical Significance legend: *p<.05, **p<.01, ***p<.001*

Cohen's d legend: no effect<.2, .2<=small<.5, .5<=medium8, large>=.8

THEMES FROM THE FOCUS GROUPS

With technical assistance from WestEd (a non-partisan, non-profit, research agency), CDE Educator Effectiveness staff conducted 19 focus groups in April 2018, gathering feedback from 82 teachers, 19 principals, and 17 district leaders involved with the pilot. WestEd transcribed the focus group recordings and analyzed all transcripts to provide an unbiased lens through which the data were interpreted. Focus group questions inquired about the impact of streamlining the standards and elements, perceptions about the scoring changes, value of the SMES for teachers, and lessons learned. Highlights from the focus groups are presented here[4].

Many of the focus group findings support the feedback obtained in the surveys and provide more detail about those perceptions. In some cases, the focus groups provided new information that the surveys did not surface. Focus group highlights included the following:

- There was a strong consensus that streamlining the teacher rubric was a beneficial change leading to a more efficient use of evaluation time and a more authentic evaluation process.
    - Educators specifically mentioned having more time in the evaluation process for shared understandings and more focused dialogue.
    - Principals reported that using the new rubric enabled more coherent goal setting because they were more able to review self-assessments prior to observations, identify trends, and engage in more meaningful conversations as part of the evaluation process.
- There was strong support for the new rubric's combining elements of professional reflection and leadership into a single standard, *Quality Standard IV – Teachers demonstrate professionalism through ethical conduct, reflection, and leadership*.
- Multiple teachers, principal/evaluators, and administrators expressed appreciation for the scoring shifts that they expected to shift some teachers from the highly effective category to the effective category. There was some concern about difficult conversations that might arise as a result, although some pointed out that difficult conversations can have positive results.
- One area that surfaced as a persistent concern across focus groups relates to local capacity for implementation of the evaluation system. This arose in conversations about rater reliability, time burdens, and capitalizing on evaluation information to inform professional development offerings, among others. As one way to ameliorate local capacity challenges, many pilot educators made recommendations for targeted areas of support that might be offered to the field by the Educator Effectiveness staff. Recommendations included "over-communication" from CDE to the field regarding the fundamental purposes of the teacher evaluation process and the rubric changes, along with suggestions for specific actions that could be taken locally to strengthen teacher ownership during the evaluation process and a request for more variation in the practical examples CDE includes in the *CDE Resource Guide*.
- Multiple respondents also recommended local leeway to tailor evaluation frequency based on teachers' previous evaluation rating or freedom for local identification of a single standard of focus for a time period, although there was recognition that this is inconsistent with current law.

[4] More detail about the focus groups conducted with the support of WestEd in April 2018 is available upon request. Please contact EducatorEffectiveness@cde.state.co.us.

Where can I learn more?

The Revised Colorado State Model Evaluation System for teachers is being offered as an option for districts and BOCES to use to comply with Senate Bill 10 191. The educator effectiveness team at CDE is releasing updated Users' and Resource Guides for the revised system.

- CDE State Model Evaluation System Pilot: http://www.cde.state.co.us/educatoreffectiveness/smes-revisions
- CDE Educator Effectiveness website: www.cde.state.co.us/EducatorEffectiveness
- Colorado State Model Evaluation System: www.cde.state.co.us/EducatorEffectiveness/StateModelEvaluationSystem