State of Idaho Department of Education: Teacher Evaluation Desk Review Report

Submitted by:
McREL International
4601 DTC Blvd., Suite 500
Denver, CO 80237-2596
P: 303.632.5631
www.mcrel.org

July 11, 2016

Table of Contents

EXECUTIVE SUMMARY

Introduction

The purpose of this project was to conduct an independent review of teacher evaluations for the State of Idaho Department of Education (SDE) to assess whether the processes and outcomes of the evaluations across the state are aligned with the goals of the SDE for the state of Idaho teacher evaluation initiative. McREL understands that large variation and/or incomplete teacher evaluation processes can lead to difficulties with decision-making for improvement and policy decisions. Therefore, this desk review of a random sample of teacher evaluations will help the SDE understand if their envisioned processes are being realized consistently across the state and whether further guidance or reform is needed for systematic improvement in the future.

Methods

At the outset of the project, McREL communicated with SDE personnel to gain knowledge of the requirements and recommendations for teacher evaluation provided by the state department to school districts and understand key interests and concerns of SDE regarding the evaluation processes. Once the initial desk review plan was finalized, McREL received from SDE 225 teacher evaluations randomly selected from 53 districts. Two lead researchers from McREL developed a database shell with variables corresponding to data elements requested by SDE as shown in Table 2 of the main report. The database contained several variables focusing on 5 major elements – 1) individualized learning plans (IPLPs), 2) classroom observations, 3) summative evaluations/ratings, 4) additional measures to inform professional practice and student achievement, and 5) written policy.

Seven experienced McREL evaluators were chosen to complete the desk reviews. Prior to assigning evaluations to individual raters, the team completed a rigorous calibration process consisting of meetings to clarify constructs and rating indicators. The team then participated in two separate rating exercises to ensure rater consistency – 1) a team rating exercise on two randomly selected evaluations, and 2) an independent rating exercise of 3 additional evaluations by all seven raters to compute statistics for inter-rater reliability. The resulting average Kappa statistic on the three evaluations for seven independent raters was .91, which is an almost perfect level of agreement (Landis & Koch, 1977). Following the calibration process, the remaining 220 evaluations, were assigned randomly and evenly to the seven evaluators by district.

Data Analysis

Data analysis consisted of calculating the number (n) and percentages of evaluations that contained evidence related to criteria specified by the SDE as shown in Table 2 – IPLPs, documented observations, summative ratings, additional measures of professional practice, and written policies. Of particular importance was information associated with use of the Charlotte Danielson Framework 2nd Edition (hereafter referred to as the Danielson Framework) (Danielson, 2011).

Findings

In general, the review of evaluations revealed a great deal of variation in reporting processes (i.e., forms and procedures) implemented throughout the state. Specifically, 34 percent of the evaluations did not include summative ratings but instead provided individual classroom observations. Furthermore, while 64 percent of evaluations included data from one observation, only 39 percent included data from a second observation. Reporting practices varied in that some evaluations included narrative only, while others used rubrics and ratings. When the Danielson Framework was used, it was often the case that components were omitted, added, and/or reworded from the Danielson Framework rubric adopted by the SDE.

A flow chart of major findings can be seen below. Reviewers found that most evaluations did not contain an IPLP or goals of any kind (55 percent). Of those evaluations that did include an IPLP, most included either one (44 percent) or two (33 percent) goals. Reviewers also noted that most (61 percent) of the goals were aligned to the Danielson Framework.

Of the 142 evaluations that included a first observation, 83 percent were based on the Danielson Framework; however only 19 percent of those included all 22 components. Of the 88 evaluations that included a second observation, 75 percent used the Danielson Framework; however only 26 percent of those included all 22 components. The 149 evaluations that included a summative form showed more consistent use of the Danielson Framework – 80 percent used the Danielson Framework, 94 percent of which included all 22 components. Furthermore, the majority (61 percent) of evaluations that included a summative form provided evidence of a 67/33 percent weighting of professional practice and student achievement.

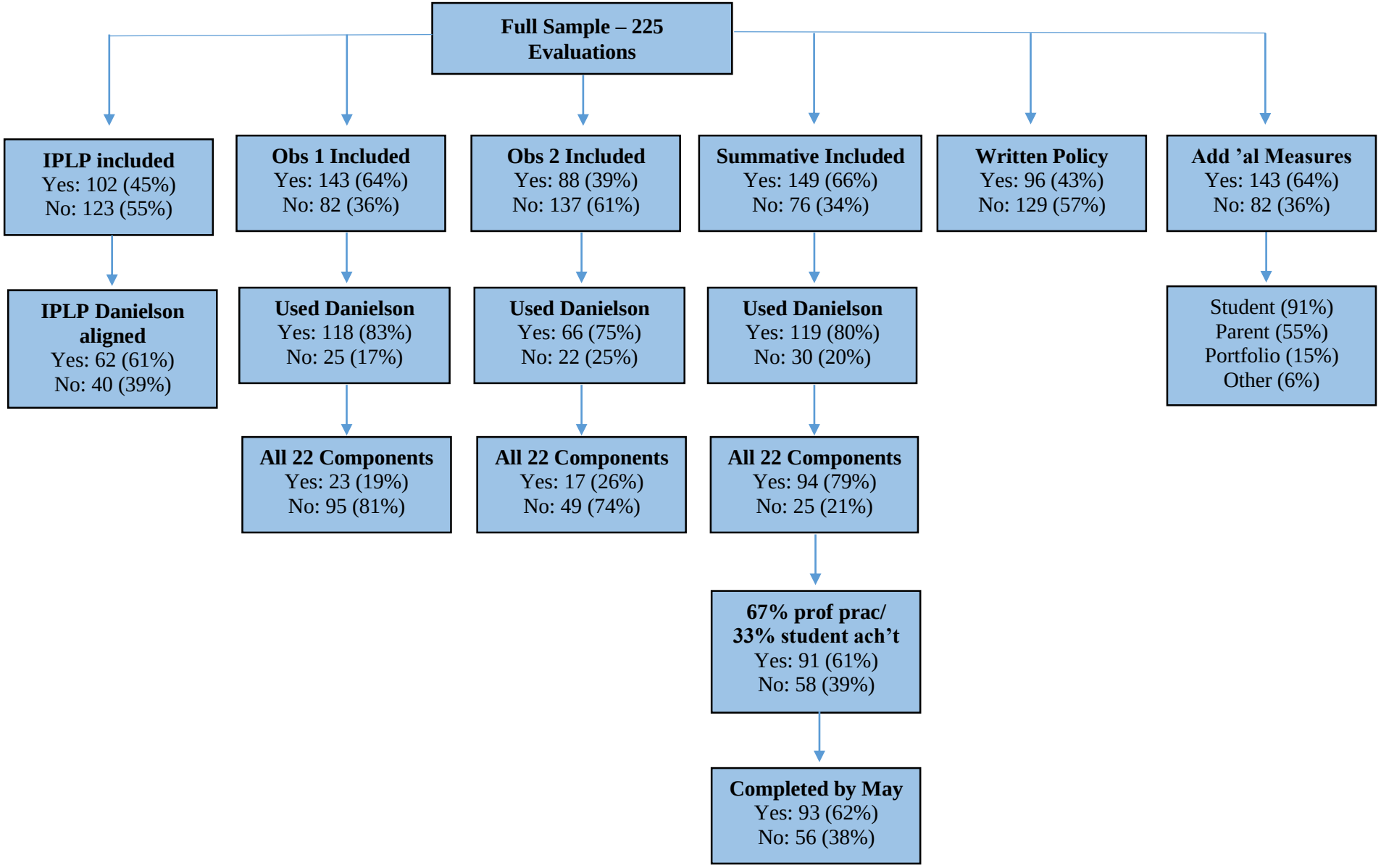
Reviewers also found that most (72 percent) of the summative forms included the performance scale of *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished*. In general, teachers were rated highly on summary rating forms, with most (between 69 and 88 percent) receiving a score of *Proficient* across the 22 components. Further, the average total score on the summary forms was a score of 3 or *Proficient*.

For those evaluations that included additional measures, reviewers found that many evaluators included student achievement (54 percent), while fewer used parent (33 percent), portfolio (9 percent), or other (4 percent) measures. For evaluations that included student measures, reviewers found that the category of "other" measures (those that did not not fall into the SDE list of possible measures) were most frequently used – examples include student survey data and student attendance. The second most frequently used measure was the Idaho Standards Achievement Tests (ISAT). Reviewers also found over half of the evaluations contained no policy information (57 percent).

Notably, only three (1 percent) of 225 evaluations contained all of the following criteria prescribed by SDE:

- Two observation time points
- Summary rating based on all 22 Danielson Framework components
- Use of the Danielson Framework rating scale (*Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished*)
- Summative rating weighting of 67 percent professional practice and 33 percent student achievement
- An overall summary rating score
- A completion date of May 1st or earlier

Flow chart of major review findings



Conclusions and Recommendations

The findings suggest a need for greater focus on consistency and adherence to key components of the evaluation system. Inconsistent implementation suggests that some districts either selected not to follow the prescribed process or lacked sufficient understanding of the system. It is recommended that the following steps be taken to avoid such implementation inconsistencies:

- Align the process of teacher evaluation to relevant policies at the state and district level to eliminate any potential conflict among policy, process, practices and procedures used to support and evaluate teachers.
- Ensure that all teachers, teacher supervisors, and central office leaders receive training on the process.
- Annually communicate to all teachers, teacher supervisors, and central office leaders the teacher evaluation process. Be specific and detailed about the roles and responsibilities of each stakeholder in order to maximize the benefit.
- Monitor and track adherence to the process to ensure consistent application.

Sound and informative teacher evaluation systems should meet the basic expectations of 1) a focus on growth and development of teachers and 2) adherence to policy. To this end, McREL recommends the following action items to ensure that the application of the evaluation system is in accordance with its ultimate intentions.

- Be sure that all teachers, supervisors and professional development staff are clear on the expectations for using and how to use the teacher evaluation rubric.
- Focus efforts on improving fidelity of performance monitoring. Questions still exist as to whether teacher supervisors know and understand what to look for and how to provide feedback to teachers based on the teacher evaluation process.
- Identify opportunities to train teachers, supervisors, and professional development staff to connect evaluation protocols to the adopted models of teacher practice.
- Be clear about the purpose of goal setting, and determine exactly how goal attainment may be used as part of the overall evaluation of teachers.
- Provide all districts with a definition of educator effectiveness that includes exactly what is expected and what measures may be used to determine an overall teacher performance score.

Introduction

McREL International (McREL) partnered with the Idaho State Department of Education (SDE) to assess a sample of teacher evaluations from Idaho public school districts. The purpose of this project was to review and document evaluation components and criteria used to evaluate teachers throughout the state. Large variation across the state in teacher evaluation processes can lead to lack of trust in the system and unreliable data. A careful review of the current processes will inform decisions for providing guidance to districts with the ultimate goal of producing a reliable state-wide system on which to base educational improvement and policy decisions.

Method

Data Source

The McREL review team obtained a random sample of 225 teacher evaluations across 53 Idaho districts from the SDE. The number of teacher evaluations per district ranged from one evaluation to 36 evaluations. Table 1 provides a list of participating Idaho school districts and the number of evaluations submitted for this study. Although the majority of evaluations received were in PDF files, teacher evaluations were also received in Word documents, Excel sheets, and TIF files.

Table 1. Districts included in the evaluation review

District Number	District Name	Number of Evaluations
56789	Bear Lake County District	3
56792	Boise Independent District	36
56793	Bonneville Joint District	13
56794	Buhl Joint District	2
56796	Caldwell District	4
56797	Cassia County Joint District	2
56798	Coeur D'alene Charter Academy District	2
56799	Coeur D'alene District	6
56802	Filer District	4
56803	Firth District	4
56805	Fruitland District	1
56806	Glenns Ferry Joint District	2
56807	Grace Joint District	2
56808	Hagerman Joint District	2
56810	Homedale Joint District	2
56811	Idaho College & Career Readiness Academy	2
56814	Idaho Falls District	2
56815	Idaho Stem Academy District	2
56816	Idaho Virtual Academy	2
56817	Jefferson County Jt District	4

District Number	District Name	Number of Evaluations
56818	Jerome Joint District	4
56819	Joint School District No. 2	10
56821	Kendrick Joint District	4
56822	Kuna Joint District	4
56823	Lakeland District	2
56824	Madison District	4
56825	Marsh Valley Joint District	2
56826	Marsing Joint District	2
56827	Middleton District	1
56828	Minidoka County Joint District	2
56829	Monticello Montessori Charter School	2
56830	Moscow District	4
56831	Mountain Home District	4
56833	Nampa School District	14
56835	North Idaho Stem Charter Academy District	2
56838	Payette Joint District	4
56839	Pocatello District	16
56840	Post Falls District	6
56841	Preston Joint District	2
56842	Ririe Joint District	2
56843	Rolling Hills Charter School	2
56846	Shoshone Joint District	2
56847	Snake River District	4
56848	Sugar-Salem Joint District	3
56849	Teton County District	5
56850	Thomas Jefferson Charter	2
56851	Twin Falls District	8
56852	Vallivue School District	2
56853	Victory Charter School	2
56854	Vision Charter School	2
56856	White Pine Charter School	2
56857	Whitepine Joint School District	1
56858	Wilder District	2
TOTAL		**225**

Review Process

Seven experienced evaluators from McREL were responsible for reviewing the teacher evaluation documents. The criteria for reviewing the evaluation documents were specified by the SDE and included the major elements of IPLPs, documented observations, summative ratings, additional measures of professional practice and student achievement, and written policy (see Table 2). The McREL review team individually examined each teacher evaluation, carefully checking for any and all teacher evaluation components of priority to SDE.

Table 2. SDE evaluation criteria

Items Reviewed	Data Element
What are the components that were on the Individual Professional Learning Plan (IPLP)?	List components
Does the professional practice portion include all 22 components of the Charlotte Danielson Framework (2nd ed.)?	Yes/No
Record the levels of performance for each component?	1, 2, 3, 4 for each component.
What are the dates of the two documented observations?	Dates
Which additional measure(s) was included to inform professional practice?	• Student Input • Parent Input • Portfolio • None
Which measures were used for student achievement?	• ISAT • Student learning objectives • Formative assessments • Teacher-constructed assessments of student growth • Pre-and post-tests • Performance based assessments • Idaho Reading Indicator • College entrance exams such as PSAT, SAT and ACT • District adopted assessment • End-of-course exams • Advance placement exams • None
What is the summative rating?	Summative Rating
Does the summative rating include combining professional practice (67%) and student achievement (33%)?	Yes/No
What is the date of the summative evaluation?	Date
Was it completed by May 1st?	Yes/No
Is there a written evaluation policy?	Yes/No

Reliability of Reviewers

To ensure consistency among reviewers, all seven McREL reviewers participated in a rater-calibration process which consisted of discussing the SDE criteria as a group, reviewing the initial coding system developed by the team leads, and participating in a calibration activity to help ensure inter-rater reliability. For this process, two evaluations were randomly selected and provided to all seven staff to review and code, once independently and again as a team. The team review process allowed the staff to discuss their assessments, clarify understanding of constructs, and rectify coding discrepancies.

To further ensure the team's ability to reliably assess the teacher evaluations, team members were provided three additional evaluations chosen at random to code independently. Following those evaluations, the Fleiss-Kappa statistic was applied to each of the variables to assess inter-rater reliability across multiple raters (Fleiss, 1981) using the MAGREE macro in SAS®9.2. For three evaluations rated by seven raters, the average Kappa statistic across all of the indicators was .91, which is considered a high level of agreement (Landis & Koch, 1977). Upon completion of the rater-calibration exercise, the McREL team leaders consulted with staff from the SDE to review the process, address questions and concerns, and clarify the use of additional variables as part of the overall study. Following the meeting and the confirmation from SDE staff, the remaining 220 of 225 evaluations were grouped by district and divided equally among McREL team members to review.

Throughout the rater-calibration process, a McREL consultant with expertise in teacher personnel evaluation helped reviewers arrive at consensus and guided discussions on possible variables that should be added to provide the SDE with a robust report. Further this team member aided in refinement of the variables specified by the SDE to accurately address essential questions concerning the effective use of the teacher evaluation system.

Data Analysis

Data analysis consisted of calculating the number (n) and percentages of evaluations that contained evidence related to criteria specified by SDE as shown in Table 2 – IPLPs, documented observations, summative ratings, additional measures of professional practice, and written policies.

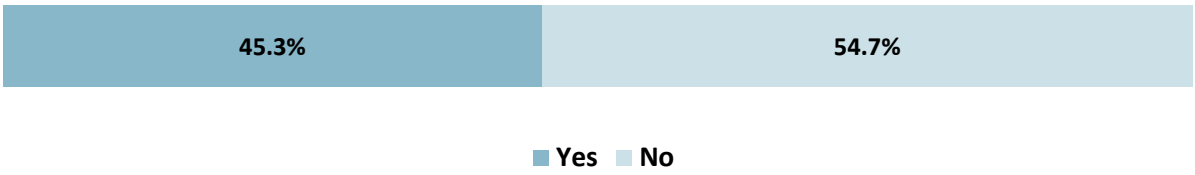
FINDINGS

The findings presented illustrate the variability across Idaho school districts in what components were included in the teacher evaluation documents and how teachers were rated on domains of the Danielson Framework. The findings are presented by major components found across the evaluations including, 1) Individualized Learning Plans (IPLPs), 2) evaluations with one observation, 3) evaluations with two observations, and 4) evaluations with a summative form.

Individualized Professional Learning Plans

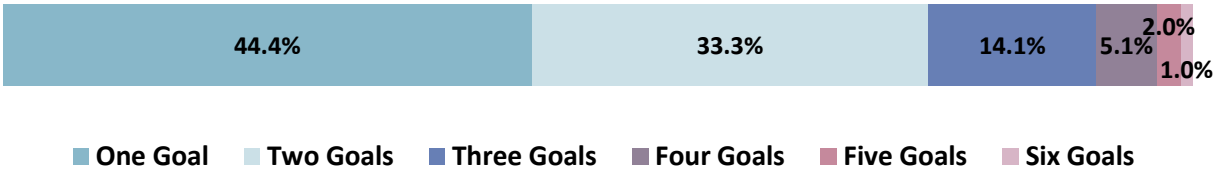
McREL reviewers examined district-provided evaluations for presence of IPLPs and key components of the IPLP including 1) whether districts provided an IPLP, 2) how many goals were set, 3) if the IPLP was aligned to the Danielson Framework, and 4) which Danielson Framework components were used to set IPLP goals. McREL reviewers found that 123 of the 225 sampled evaluations did not include an IPLP (see Figure 1).

Figure 1. Was an IPLP or personalized goal(s) included? (*n* = 225)



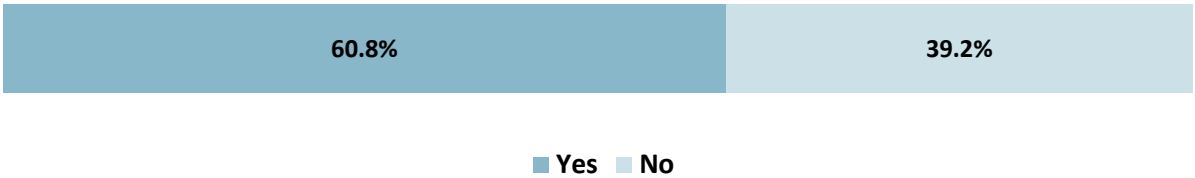
Of the 102 evaluations that included an IPLP, most (92 percent) included between one and three goals. In a few cases, reviewers found that teacher evaluations contained four or more goals (8 percent). These results are presented in Figure 2.

Figure 2. How many personalized goals were included in the IPLP? (*n* = 102)



McREL reviewers also found that of the 102 evaluations that included IPLPs, 62 (61 percent) aligned to the Danielson Framework. For evaluations with IPLPs that did not align to the Danielson Framework, teachers might have set their own goals, integrated a school- or district-wide goal, or used something else not aligned with Danielson components. These results are broken down by percentage and can be reviewed in Figure 3.

Figure 3. Was the IPLP or personalized goals aligned to Danielson components? (*n* = 102)



McREL reviewers tabulated the number of times specific Danielson Framework domains and components were used to create IPLP goals (see Figures 4 through 7). At the domain level, Domain 3 (*Instruction and Use of Assessment*) was the most frequently used domain (n = 33), whereas Domain 4 (*Professional Responsibilities*) was the least frequently used (n = 24). Across specific components, 3c (*Engaging Students in Learning)* was the most commonly used component in IPLPs (n = 12), followed by 3d (*Using Assessment in Instruction*; n = 10), 1a (*Demonstrating Knowledge of Content and Pedagogy*; n = 9), and 4c (*Communicating with Families*; n = 9). The least frequently used components were 1f (*Designing Student Assessments*), 2d (*Managing Student Behavior*), 4a (*Reflecting on Teaching*), and 4f (*Showing Professionalism)*, each of which were used in only one evaluation.

Figure 4. IPLP goal alignment to Domain 1: Planning and Preparation (n = 28)

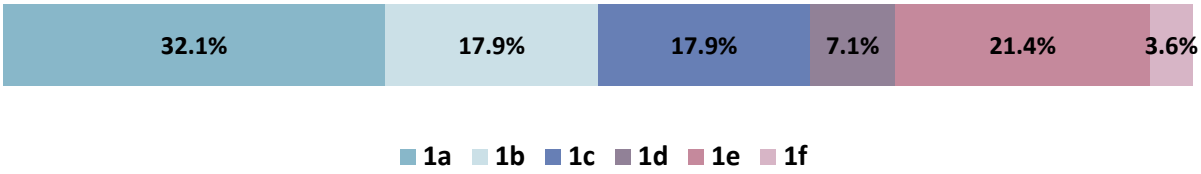


Figure 5. IPLP goal alignment to Domain 2: Learning Environment (n = 25)

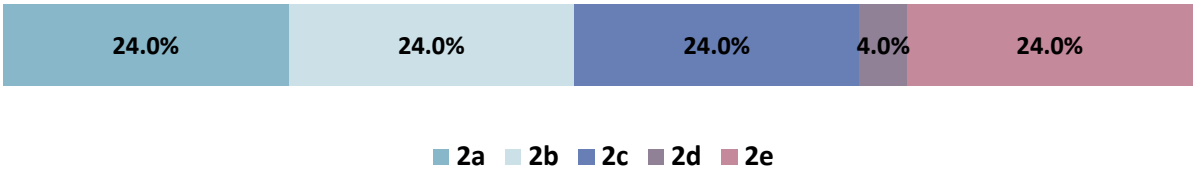


Figure 6. IPLP goal alignment to Domain 3: Instruction and Use of Assessment (n = 33)

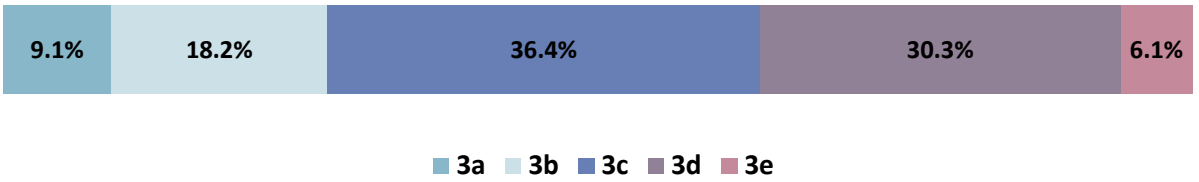
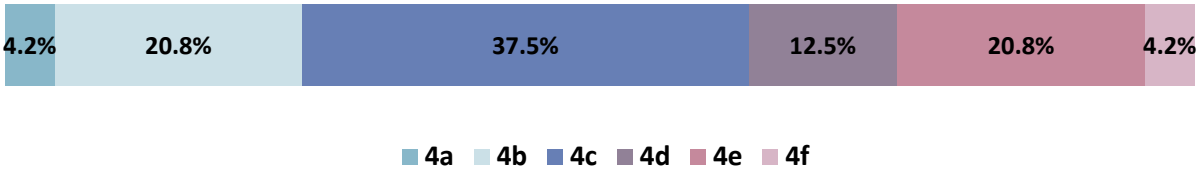


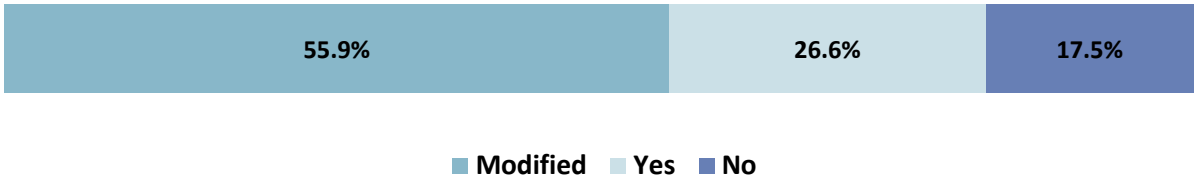
Figure 7. IPLP goal alignment to Domain 4: Professional Responsibilities (n = 24)



Evaluations with One Observation

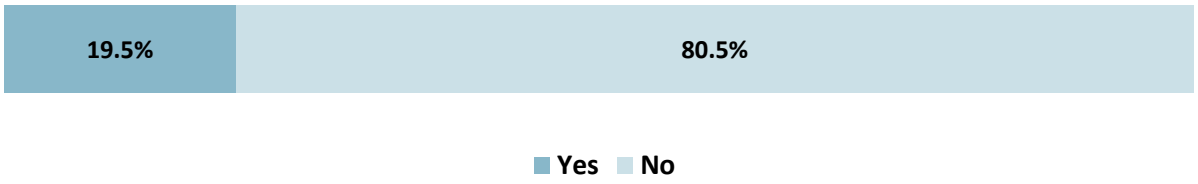
McREL found that 143 of the evaluations included at least one individual classroom observation in addition to (or in place of) a summative evaluation. To capture information from observations provided, McREL reviewers looked for observational tools and instruments that included aspects of the Danielson Framework. As shown in Figure 8, 118 of the 143 were based on the Danielson Framework or used a modified version (i.e., included customized components or omitted components).

Figure 8. Was the Danielson rubric (2nd Ed.) used? (*n* = 143)



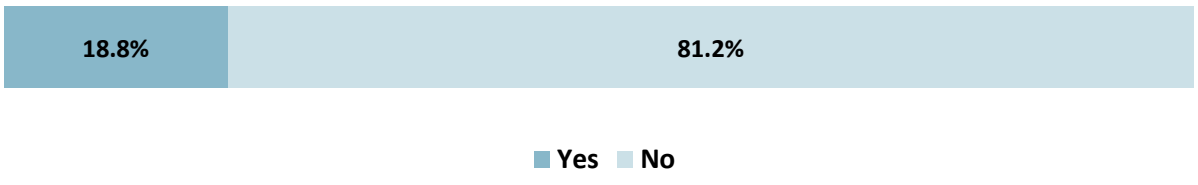
Of the 118 observations that were based on the Danielson Framework, 95 did not use all 22 Danielson components in the observation process. In some cases, only a small set of components were used. For other observations, scores for an entire domain were presented, but individual ratings on the components were not provided. These results can be reviewed in Figure 9.

Figure 9. Does observation one include all 22 components of the Charlotte Danielson Framework (2nd Ed.)? (*n* = 118)



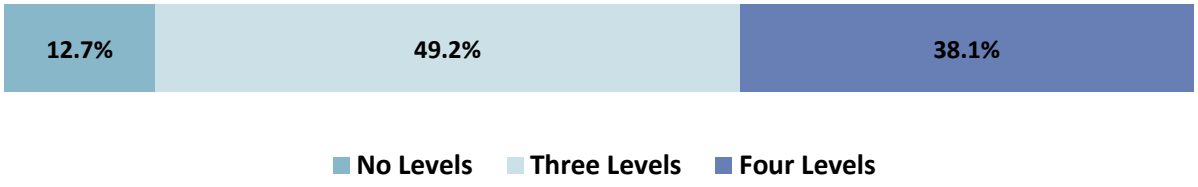
Furthermore, only 22 (19 percent) of the 118 observations that used the Danielson Framework for a first observation used the Idaho four-point rating scale of *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished* (Figure 10). McREL reviewers noted wide variation in the scales used in observations. Some districts excluded the lowest point (*Unsatisfactory*) or highest point (*Distinguished*) on the scale. Others used different scales that did not align with the SDE adopted response scale.

Figure 10. Is the 4-point Idaho adopted performance scale used? (*n* = 118)



With regard to rating scales, 58 (49 percent) used a three-point scale, 45 (38 percent) used a four-point scale, and 15 (13 percent) included no performance levels (e.g., used check boxes, or yes/no responses). The percentage of evaluations including each of these observational performance levels are presented in Figure 11.

Figure 11. How many performance levels were included in observation one? (*n* = 118)



For observations in which teachers were rated using the Danielson Framework, the vast majority received a rating of *Proficient* across all components in Domains 1 to 4. This lack of variation indicates teacher evaluators typically did not use the lower end of the scale (e.g., *Unsatisfactory* or *Basic*) and instead generally rated teachers as *Proficient* or *Distinguished*. No teachers received a rating of *Unsatisfactory* in any of the four domains. A breakdown of these percentages and the count for each component can be reviewed in Figures 12 to 15.

Figure 12. Performance ratings for Domain 1: Planning and Preparation

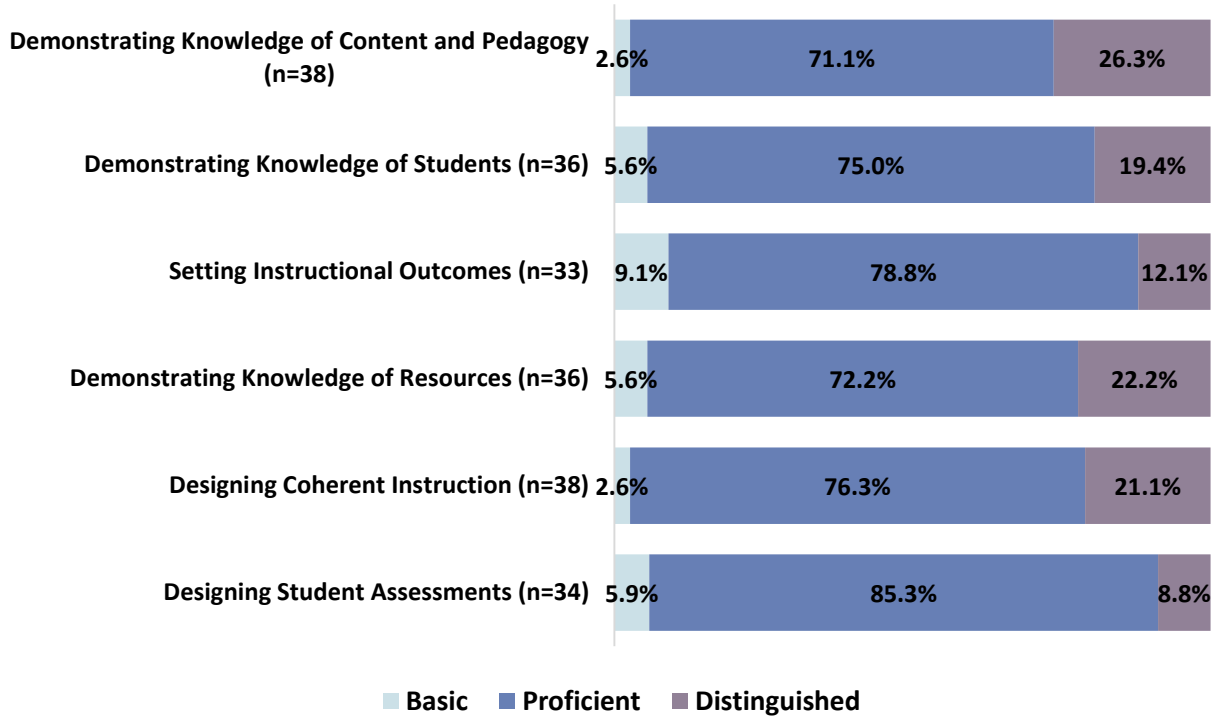


Figure 13. Performance ratings for Domain 2: Learning Environment

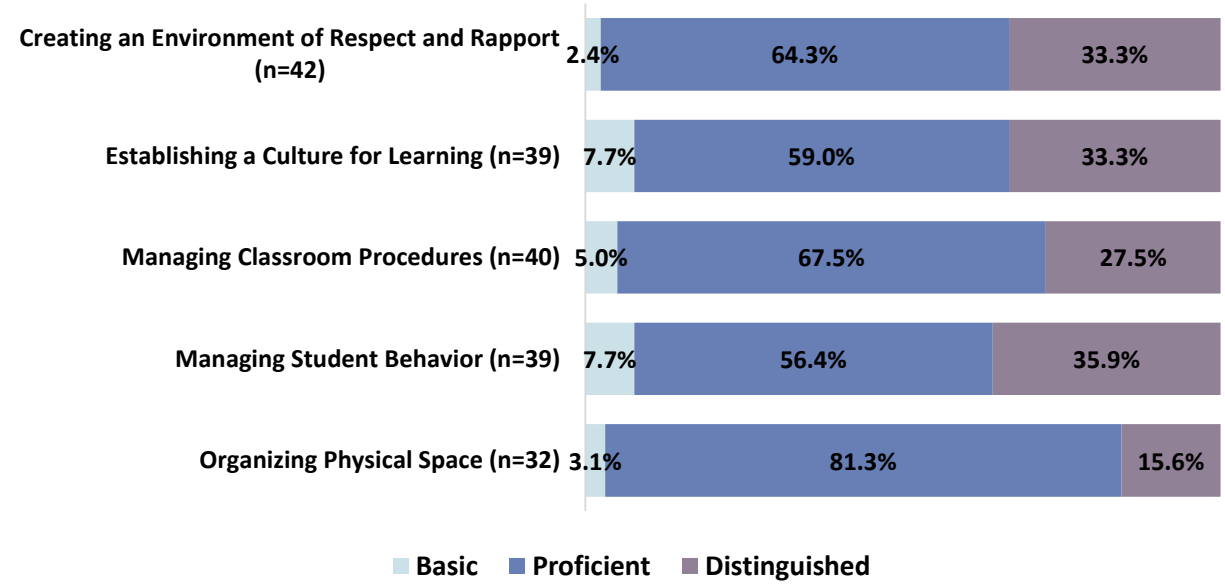


Figure 14. Performance ratings for Domain 3: Instruction and Use of Assessment

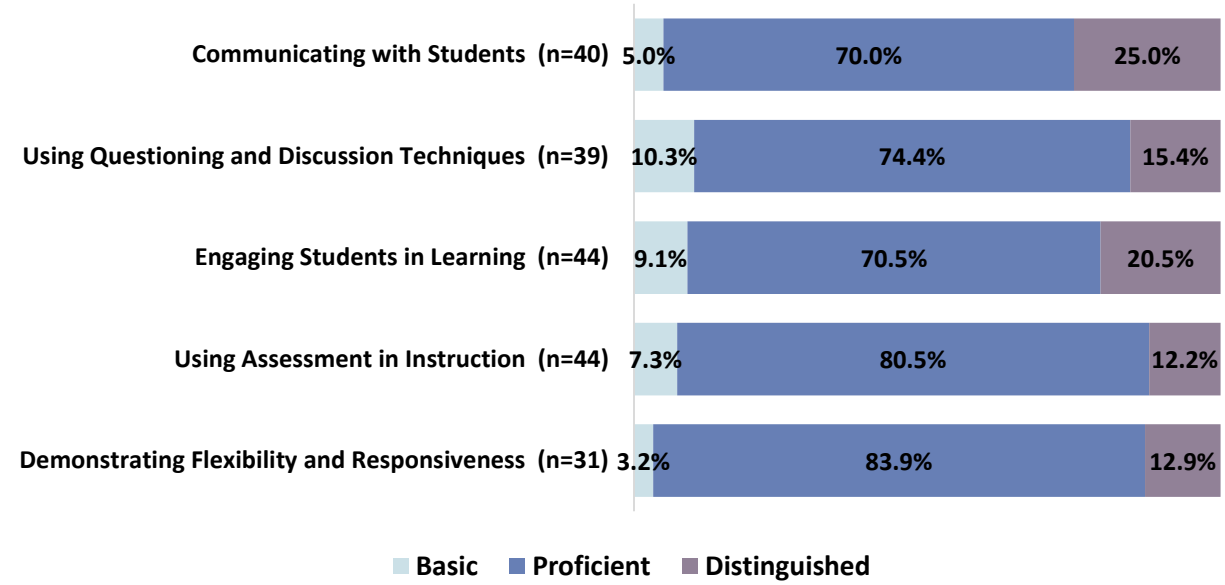
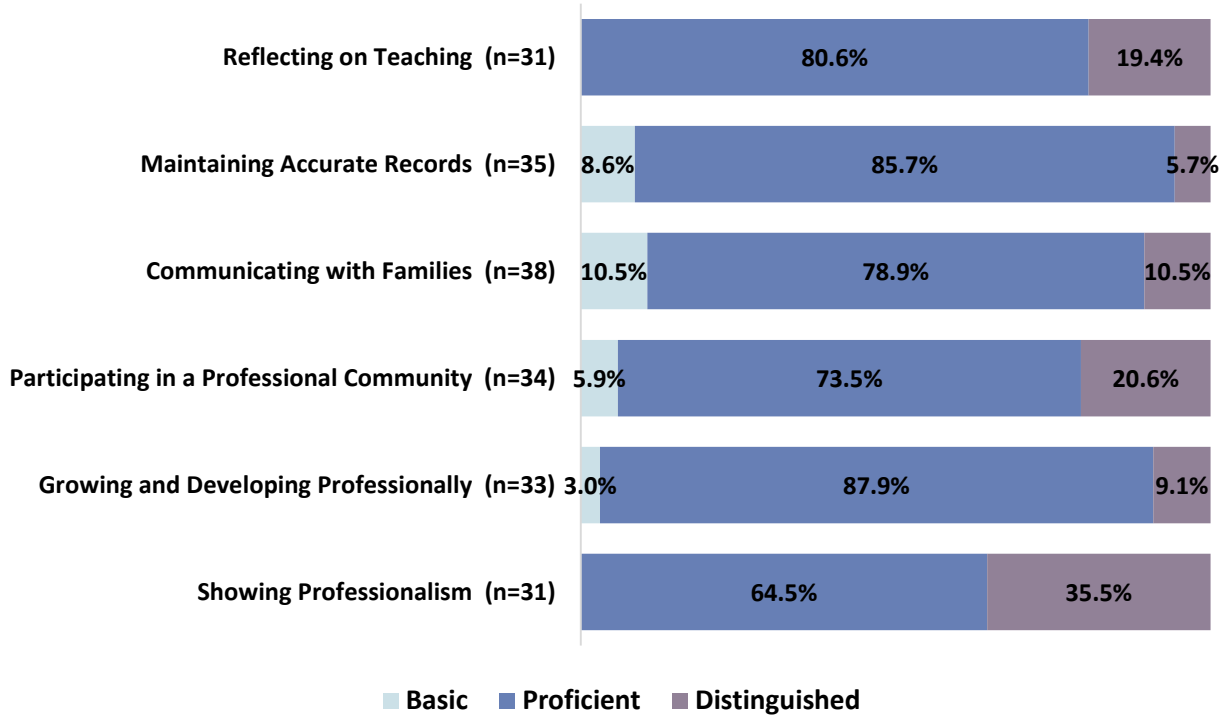


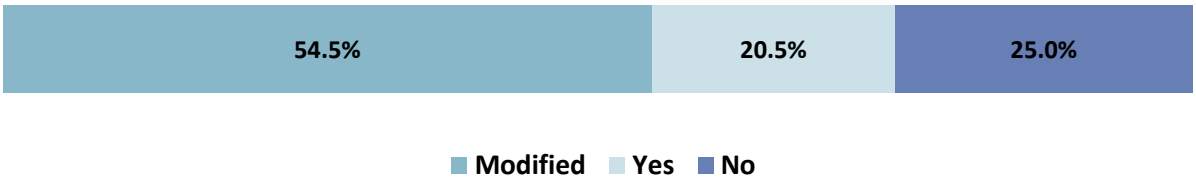
Figure 15. Performance ratings for Domain 4: Professional Responsibilities



Evaluations with Two Observations

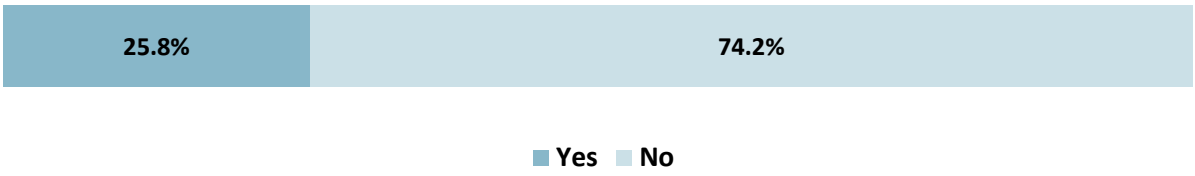
Eighty-eight of the 225 evaluations contained a second classroom observation[1]. Of those 88 evaluations, 66 were based on the Danielson Framework (*n* = 18) or used a modified version (*n* = 48) (see Figure 16).

Figure 16. Was the Danielson rubric (2nd Ed.) used? (*n* = 88)



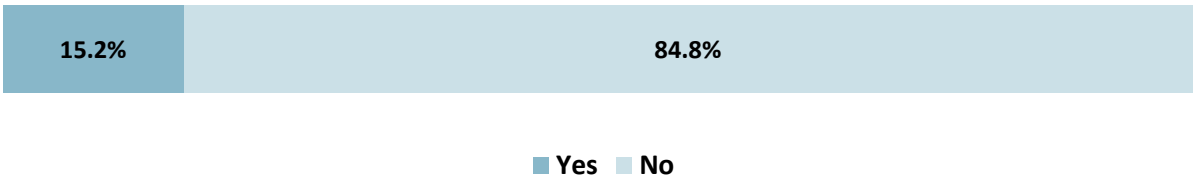
In the 66 evaluations where a second observation was included, researchers found evidence of all 22 Danielson components used in 17 of the observations (see Figure 17). Similar to the first observations, in some cases only a small set of components were used, or scores for an entire domain were presented, but individual ratings on the components were not provided.

Figure 17. Does observation two include all 22 components of the Charlotte Danielson Framework (2nd Ed.)? (*n* = 66)



Of the 66 evaluations that included a second observation, 10 used the Idaho approved rating scale of *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished* (see Figure 18).

Figure 18. Is the Idaho adopted performance scale used? (*n* = 66)



[1] Some evaluations contained more than two observations. In these cases, reviewers entered data from the first observation recorded and the last observation recorded.

McREL reviewers found that for evaluations that contained a second observation, 53 percent used 3 performance levels (e.g., *Basic*, *Proficient*, and *Distinguished*), 38 percent used four levels (e.g., *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished*), and 9 percent used no performance levels (e.g., check boxes or yes/no responses) (see Figure 19).

Figure 19. How many performance levels were included in observation two? (*n* = 66)

For second observations in which teachers were rated using the Danielson Framework, the majority received a rating of *Proficient* across the four domains. As was the case for the first observations, this lack of variation in the second observations indicates that evaluators typically did not use the lower end of the scale (e.g., *Unsatisfactory* or *Basic*) and instead generally rated teachers as *Proficient* or *Distinguished*. No teachers received a rating of *Unsatisfactory* in Domains 2 and 4. A breakdown of these percentages can be seen in Figures 20 to 23.

Figure 20. Performance ratings for Domain 1: Planning and Preparation

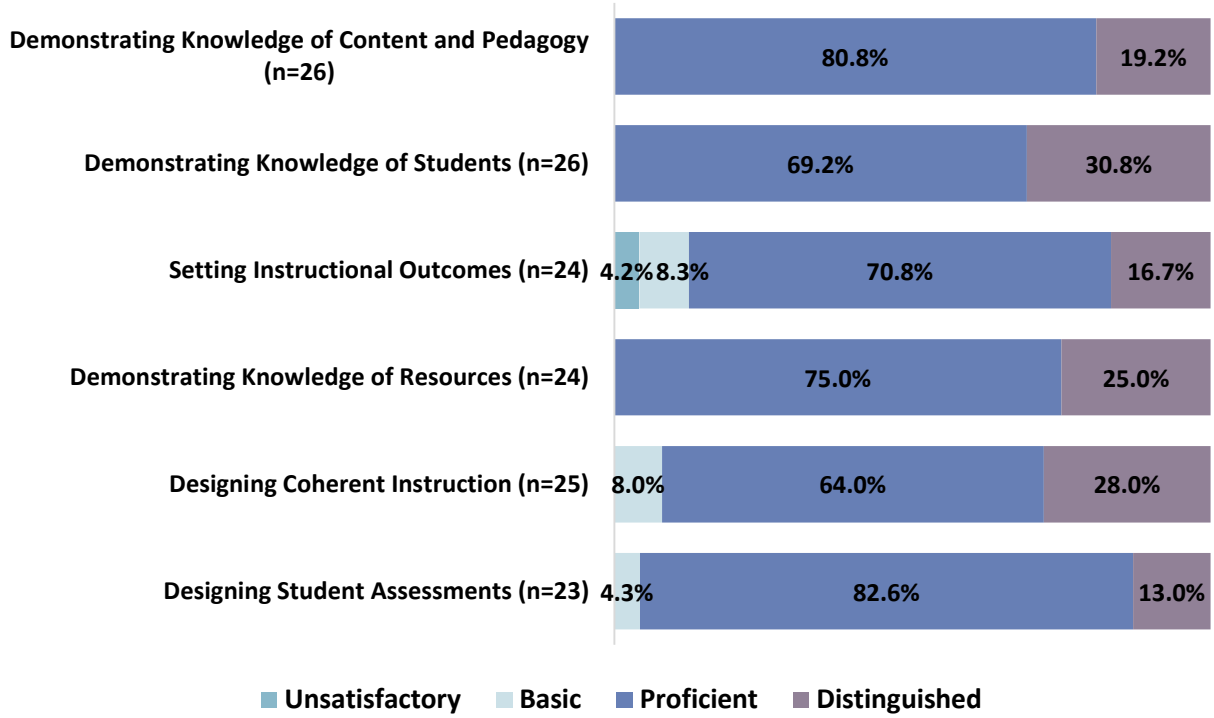


Figure 21. Performance ratings for Domain 2: Learning Environment

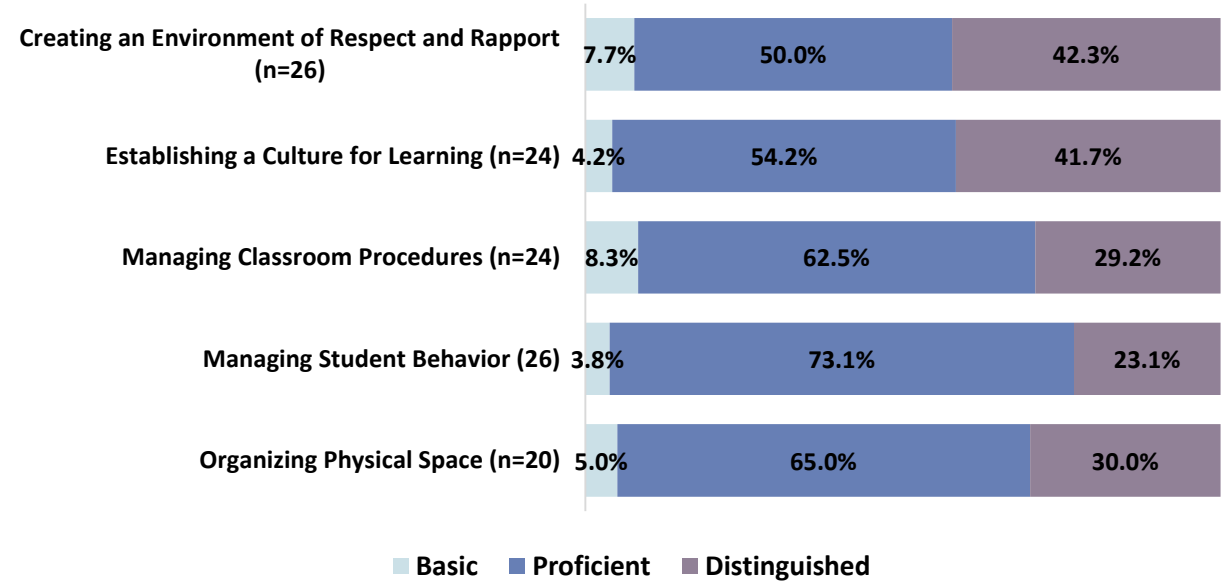


Figure 22. Performance ratings for Domain 3: Instruction and Use of Assessment

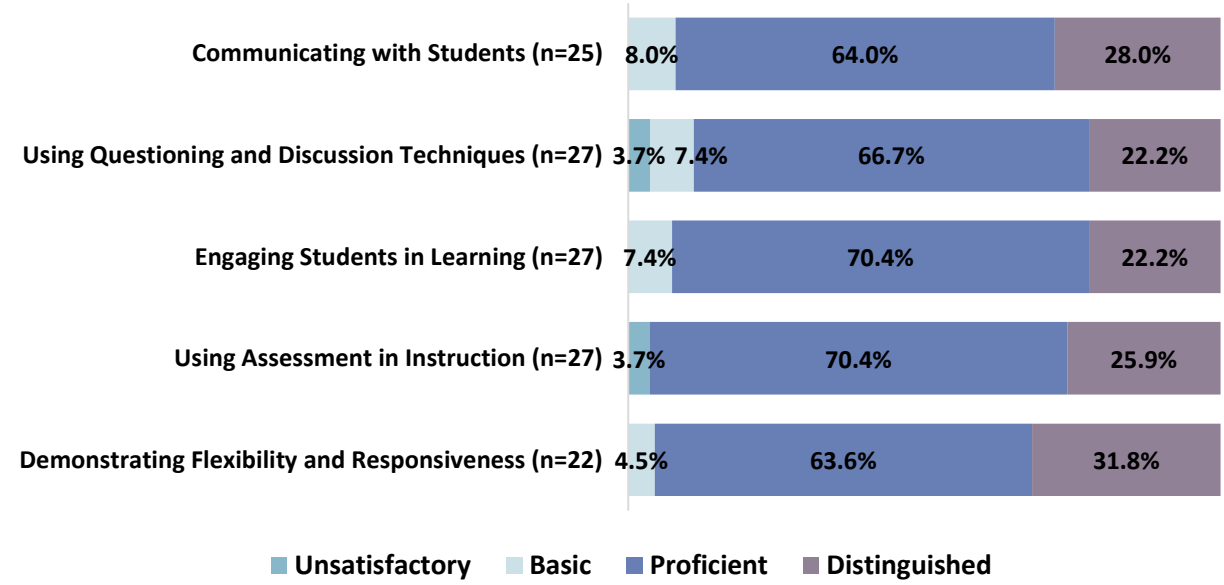
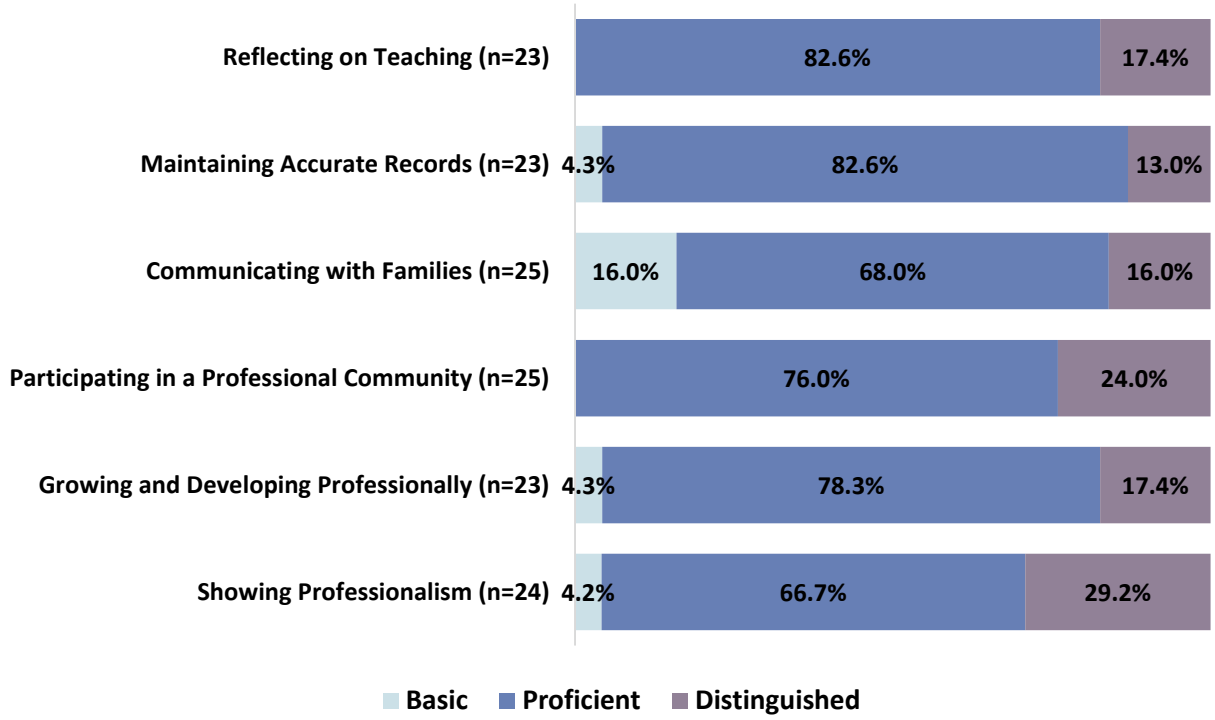


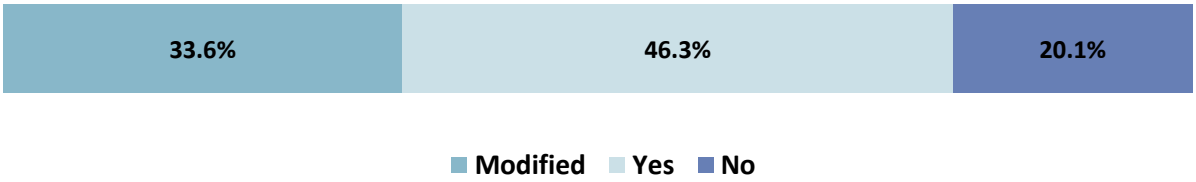
Figure 23. Performance ratings for Domain 4: Professional Responsibilities



Summative Evaluations

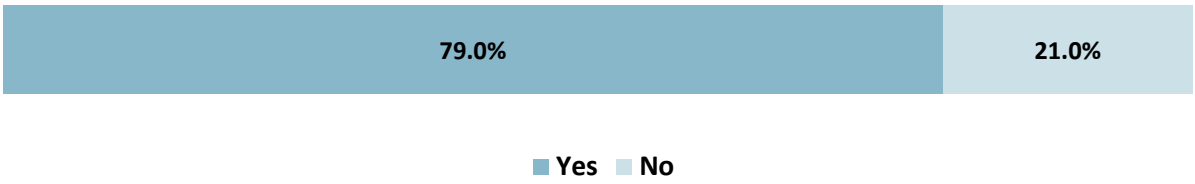
One-hundred and forty-nine (66 percent) of the 225 evaluations included a summative evaluation. For those that did, 119 (80 percent) used the Danielson Framework or a modified version. These results can be viewed in Figure 24.

Figure 24. Was the Danielson rubric (2nd Ed.) used in the summative evaluation? (*n* = 149)



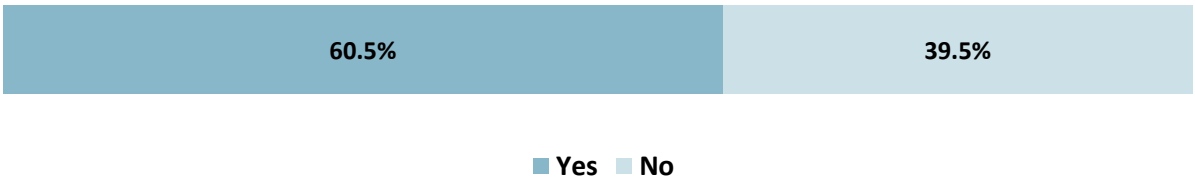
Of the 119 evaluations that included a summative evaluation based on the Danielson Framework, 94 included all 22 components across Domains 1 through 4 (see Figure 25)[2].

Figure 25. Does the summative evaluation include all 22 components of the Charlotte Danielson Framework (2nd Ed.)? (*n* = 119)



Seventy-two out of 119 evaluations that included a summative evaluation used the scale of *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished*. In some cases, however, McREL reviewers found different terms were used such as *Exceeds Expectations* rather than *Distinguished*. These findings are shown in Figure 26.

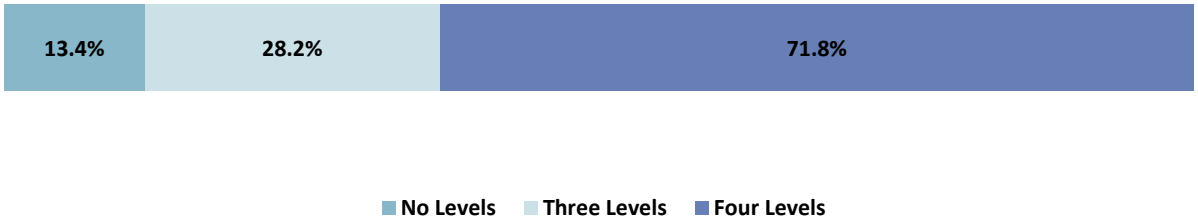
Figure 26. Is the Idaho adopted performance scale used? (*n* = 119)



[2] Some evaluations that received a "yes" for including all 22 components used additional components not specified in the SDE adopted framework.

As shown below in Figure 27, The majority of evaluations used a four-point scale (72 percent), such as *Unsatisfactory*, *Basic*, *Proficient*, and *Distinguished*. In some cases, however, reviewers found the lowest or highest level was omitted from the evaluation form, creating a three-point scale. Furthermore, 13 percent of the summative evaluation forms did not include a performance scale.

Figure 27. How many performance levels were included? (*n* = 119)



In the summative evaluations, *Proficient* was most commonly used rating, averaging 80 percent across all of the components of Domains 1 through 4. The rating of *Distinguished* was the next most common rating (with an average of 16 percent). On average, only 4 percent of teachers received a rating of *Basic*, and 1 percent a rating of *Unsatisfactory*. These results can be reviewed in Figures 28 to 31.

Figure 28. Summative performance ratings for Domain 1: Planning and Preparation

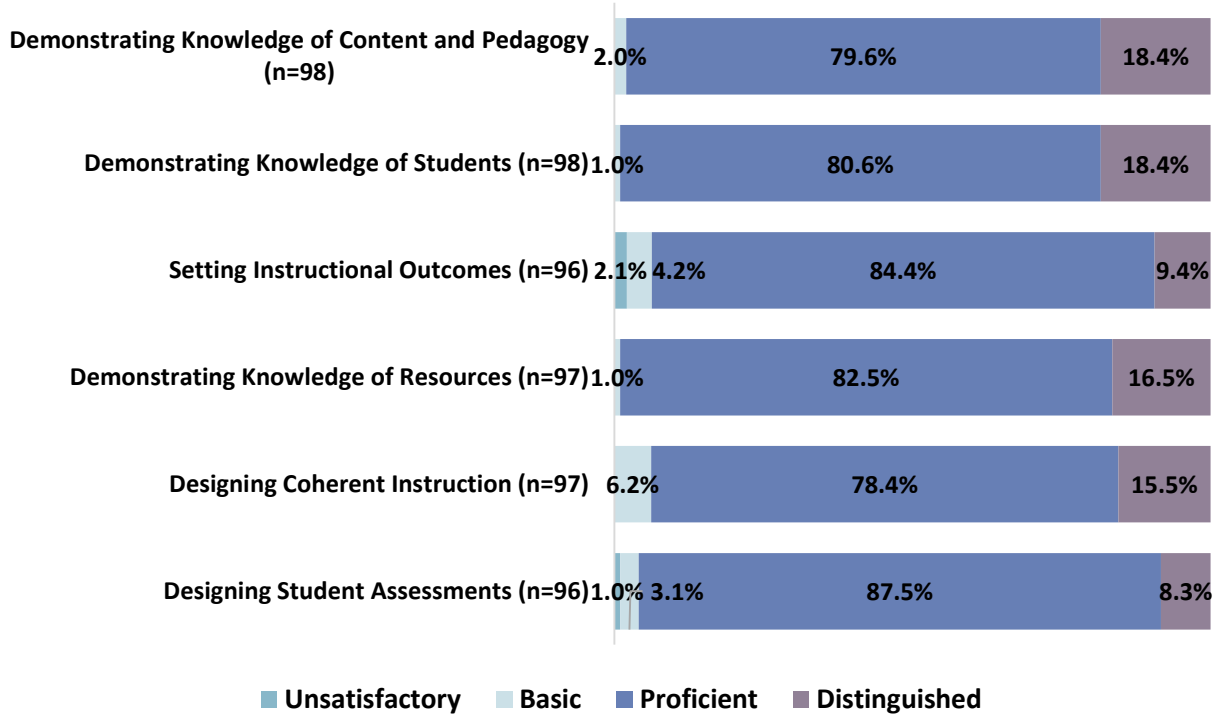


Figure 29. Summative Performance ratings for Domain 2: Learning Environment

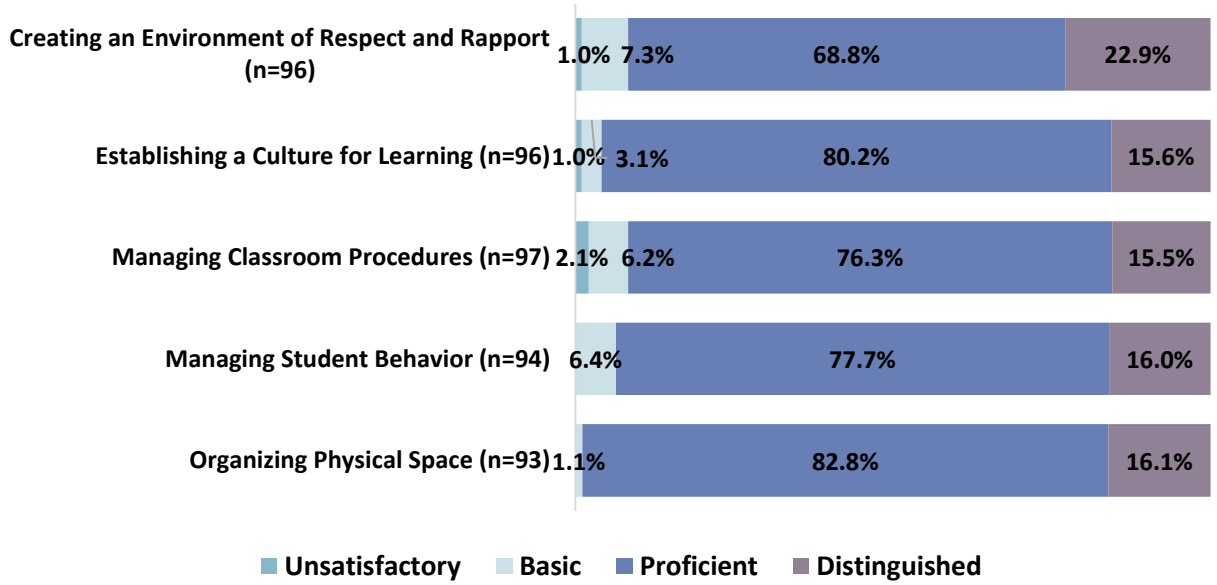


Figure 30. Summative Performance ratings for Domain 3: Instruction and Use of Assessment

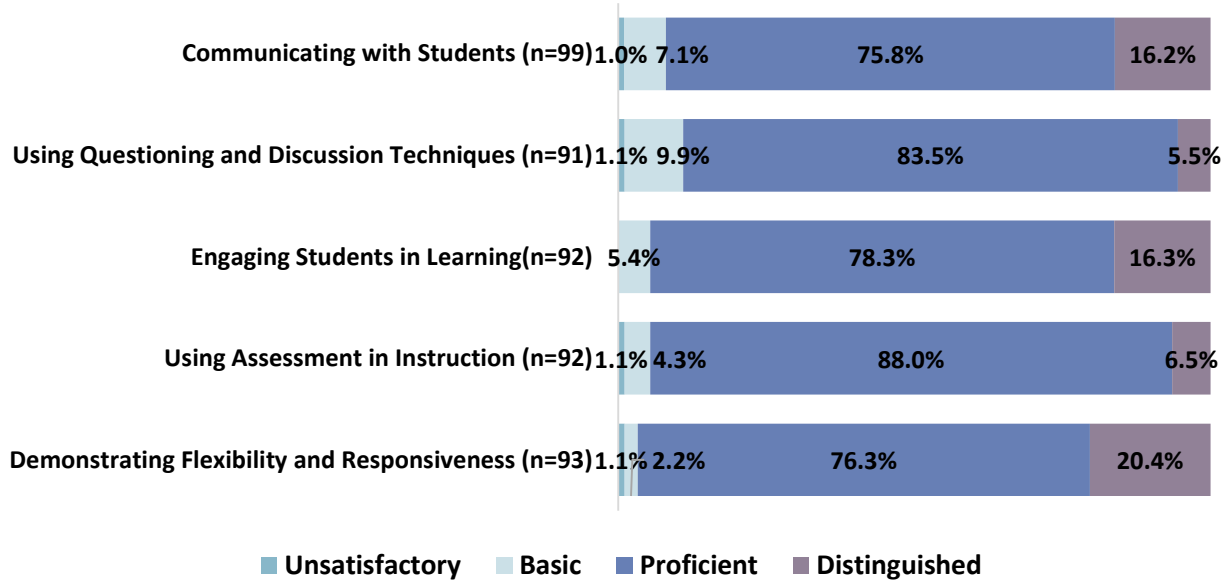
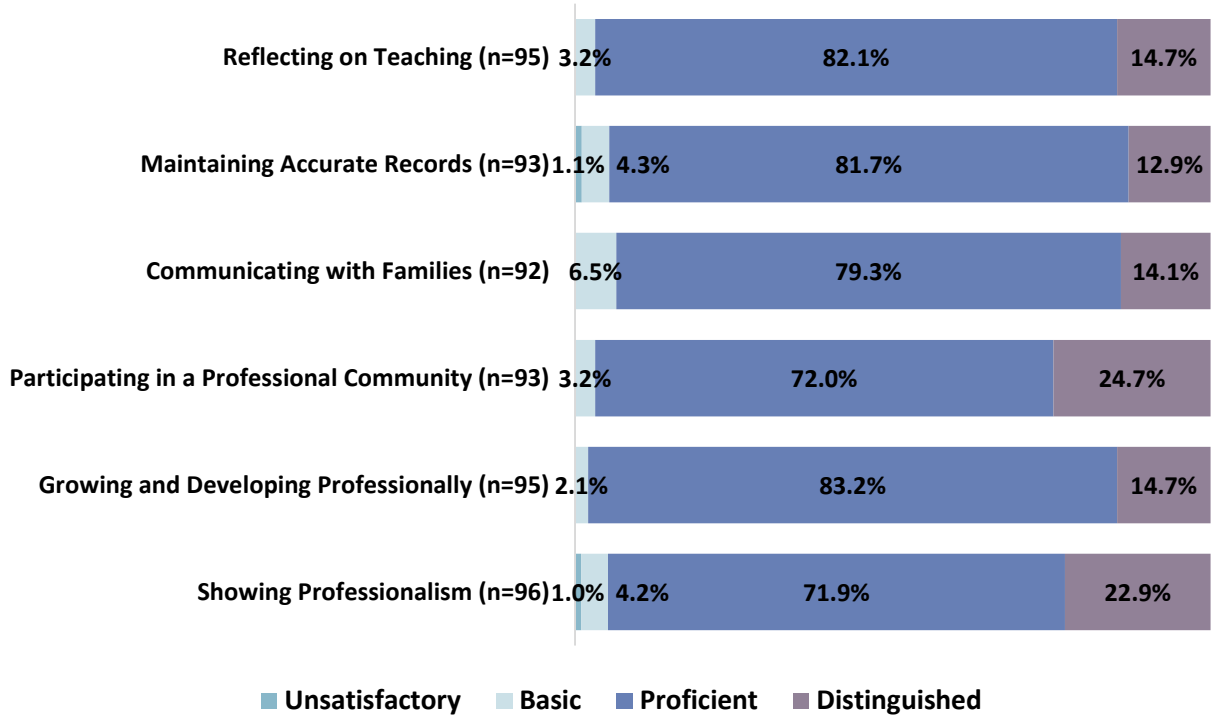
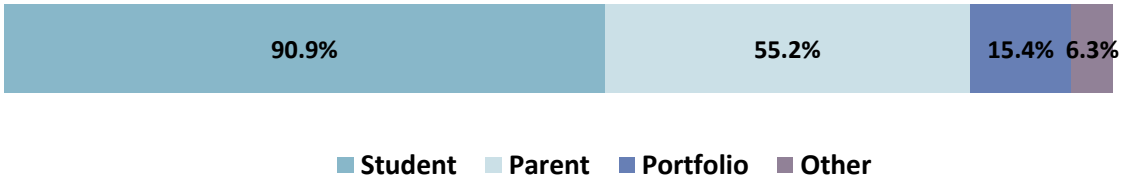


Figure 31. Summative Performance ratings for Domain 4: Professional Responsibilities



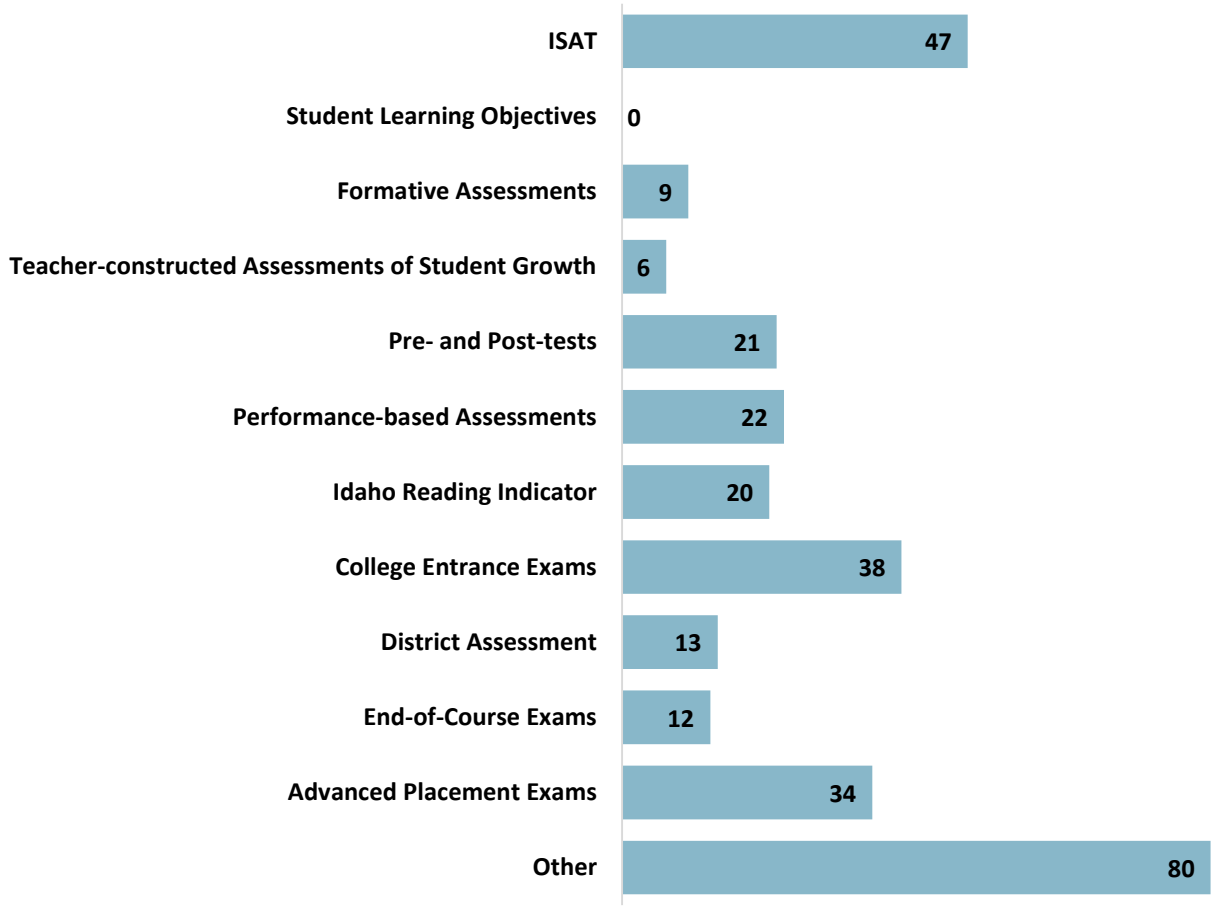
Of the 225 total evaluations selected for review, 143 used additional measures. Fifty-nine evaluations included 1 additional measure (41 percent), 74 included 2 additional measures (52 percent), and 10 included 3 additional measures (7 percent). The most commonly used additional measures were student measures included in 130 evaluations (91 percent). The next most common additional measure was parent input included in 79 evaluations (55 percent). Percentages of use across the additional measures can be seen in Figure 32.

Figure 32. Which additional measure(s) was included to inform professional practice? (*n* = 143)



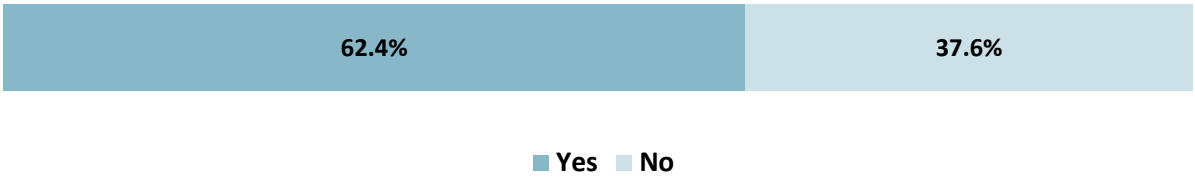
"Other" was the most common type of student measure used. The "other" measure consisted of measures not captured by the categories in Figure 33 and included measures such as student survey data, reading speed and accuracy, student portfolios, student attendance). The ISAT was the second most commonly used student measure. Counts for types of additional measures used can be reviewed in Figure 33.

Figure 33. Which measures were used for student achievement?



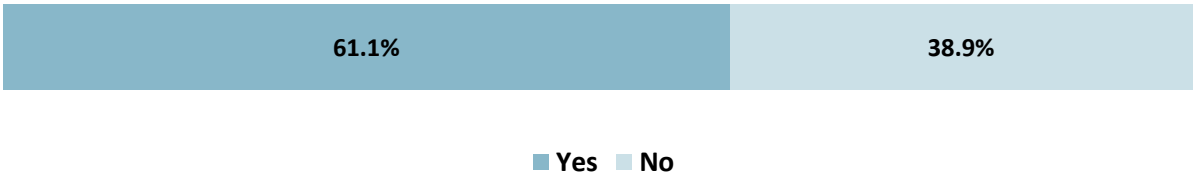
McREL reviewers found that of the 149 evaluations that included summative evaluations, most were completed by May 1st 2015. These results are presented in Figure 34.

Figure 34. Was the summative evaluation completed by May 1st? (*n* = 149)



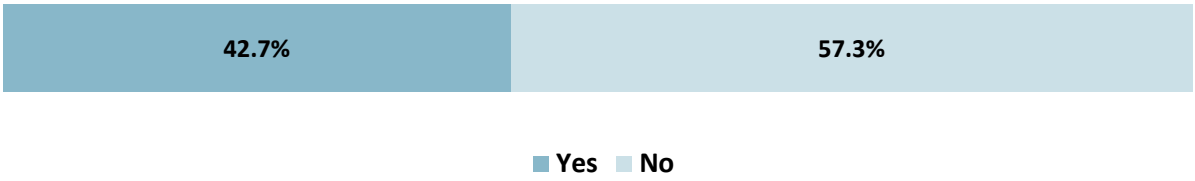
For those evaluations that included both professional practices and student achievement measures in the summative evaluation, most integrated these measures in the recommended 67 percent (professional practices) and 33 percent (student achievement) proportions (see Figure 35).

Figure 35. Does the summative rating include combining professional practice (67%) and student achievement (33%)? (*n* = 149)



McREL reviewers checked all of the evaluation documents submitted to the SDE to determine if policy information regarding the evaluation process was included. Slightly less than half of the evaluations included policy information as can be seen in Figure 36.

Figure 36. Did they include policy or procedural guidelines for the evaluation process? (*n* = 225)



In total, 87 (58 percent) of the 149 summary evaluations included a summary rating. Three ratings were excluded from our analysis as they were extreme outliers (e.g., a score over one-hundred). For the coding, McREL reviewers used the following values to code proficiency ratings, *Unsatisfactory*=1, *Basic*=2, *Proficient*=3, and *Distinguished*=4. Of the 84 ratings used in this analysis, the mean score was 2.99 with a standard deviation of 0.31. When including only those summative evaluations that used a four-point scale (51), the mean was virtually unchanged at 3.00, and a standard deviation of 0.27. For cases where summary rating forms used a three-point scale (22) the mean was 2.97 with a standard deviation of 0.11. These results suggest that in general, teachers were rated as *Proficient* most frequently across teacher evaluations that included a summative evaluation.

Conclusions and Recommendations

The selection and implementation of the Idaho teacher evaluation system reinforces a comprehensive plan and effort by the SDE to support improvements in academic achievement and outcomes for all Idaho students. Nationally, the priority on improving student achievement has focused on a number of strategies, of considerable note, the selection, implementation and effective use of systems to support and evaluate teachers and leaders.

Just in front of leadership, instructional quality is the most significant factor linked to improved student achievement (Beasley & Apthorp, 2010; Darling-Hammond, 2000; Hattie, 2008; Sanders & Rivers, 1996). To advance teaching and learning, the field has called for stronger preparatory programs, meaningful professional development resources, and refined tools to support best practices within the teaching workforce. Meeting the expectation of supporting teacher performance through evaluation and feedback, the SDE adopted the Danielson Framework and provided specific reporting requirements for schools and districts.

The current desk review was designed to examine the application and use of the current Idaho teacher evaluation system. Doing so led us to consider recommendations that address the following:

- consistency in the way the process was applied across schools and school districts;
- fidelity, or the degree to which school districts implemented the system the way it was intended to be implemented; and,
- utility, or the usefulness and benefit, to school systems across the state for improving instructional quality.

Consistent Implementation

It is clear from the findings that the Idaho teacher evaluation system was implemented inconsistently. Inconsistent application can lead to several obstacles that hinder teacher development and improved instructional quality. By contrast, the advantage of a consistently applied system is high quality feedback provided to teachers and leaders with the ultimate goal of improved practice. Individually, well implemented systems can support teacher performance and provide pathways for professional growth. Systemically, the data offers school districts with important information about the strengths and gaps in teacher performance and can assist in making decisions about system-wide teacher preparation, talent identification, and professional development.

Inconsistent implementation suggests that some districts either selected not to follow the prescribed process or held an insufficient understanding of the system. For either reason, it is recommended that the following steps be taken to avoid such implementation inconsistencies:

- Align the process of teacher evaluation to relevant policies at the state and district level to eliminate any potential conflict among policy, process, practices and procedures used to support and evaluate teachers.
- Ensure that all teachers, teacher supervisors, and central office leaders receive training on the process.
- Annually communicate to all teachers, teacher supervisors, and central office leaders the teacher evaluation process. Be specific and detailed about the roles and responsibilities of each stakeholder in order to maximize the benefit.
- Monitor and track adherence to the process to ensure consistent application.

Fidelity to key components

Report findings also suggest a lack of fidelity to some essential components expected by the system. The absence of fidelity, or lack of alignment between intended and actual use, is likely to compromise the effectiveness of the system.

Report findings demonstrate that less than half of the sampled evaluations included a meaningful and measurable Individualized Professional Learning Plan (IPLP). However, more than 60 percent of the evaluations that contained an IPLP did, in fact, align with the Danielson Framework.

In addition, 36 percent of evaluations provided no evidence of a first observation and 61 percent of evaluations provided no evidence of a second evaluation. Moreover, 34 percent of the evaluations did not include a summative form. Lastly, 39 percent of the evaluations did not use the recommended 67/33 percent weighting of professional practice/student achievement, and 38 percent were not completed by May.

One aspect of the evaluation process showed a more promising level of fidelity; specifically, of the 76 percent of evaluations that included a summative form, 80 percent used the Danielson Framework, and of those, 79 percent used all 22 components.

Finally, this report suggests that a variety of measures were used as additional measures in the evaluations. If the expectation of determining teacher performance is partly based on additional measures beyond in-class observations, then those additional measures should be standardized to ensure the process is conducted with equity. While the process of evaluation is formative (i.e., uses real-time evidence to improve and shape the quality of teaching), there must also be a summative component, which utilizes formative data combined with student achievement measures to "sum up" the overall performance or status of a teacher.

Teacher evaluation typically meets two basic expectations – (1) growth and development of teachers and (2) compliance to policy. To maximize the extent to which fidelity in each step of the process is followed to meet the basic expectations, McREL suggests the following:

- Be sure that all teachers, supervisors and professional development staff are clear on the expectations for using and how to use the teacher evaluation rubric.
- Focus efforts on improving fidelity of performance monitoring. Questions still exist as to whether teacher supervisors know and understand what to look for and how to provide feedback to teachers based on the teacher evaluation process.
- Identify opportunities to train teachers, supervisors, and professional development staff to connect evaluation protocols to the adopted models of teacher practice.
- Be clear about the purpose of goal setting, and determine exactly how goal attainment may be used as part of the overall evaluation of teachers.
- Provide all districts with a definition of educator effectiveness that includes exactly what is expected and what measures may be used to determine an overall teacher performance score.

Utility of the system

While the current evaluation framework and supporting system is effectively utilized by many other school systems across the nation, it was important for this study to examine the utility of how the system provides the best opportunity for improving instructional quality across Idaho school districts. Often utility of a system or practice is predicated on the fidelity – the way in which the system or practice is exemplified and the consistency in which the system or practice is applied. The findings of the current review suggest a compromise to both. Thus, the potential for educators to get the expected outcomes from using the system to meet performance improvement expectations is significantly diminished.

Arguably, the difference or variance in teacher practice can, and often does, result in disparate opportunities for student success. As reported in the Widget Effect (2009), teachers vary widely in their effectiveness, and as a result, such variability in practice often has negative effects on student achievement. Implied by most current systems to evaluate teachers, there should be an increase in teacher quality and reduction in variability with proper implementation. This is not to suggest taking away teacher autonomy; rather, systems should make sure that teachers can execute basic "proficient" teaching knowledge and skills before turning to creativity and innovation as a primary strategy. Over time, if one expects to see improved results from increasing teacher quality and reducing the variability in that quality, then it is evident that effective use of this or any system of support and evaluation is more likely and sustainable over time. As such, it is recommended that SDE consider the following:

- Develop local and state capacity to deliver professional development and support implementation. This will aid in avoiding the effects of personnel turnover with the knowledge and skill to provide professional development and support the effective use of the system.
- Require all educators new to the system to receive training from an authorized (local or state) trainer. The state may consider utilizing existing professional organizations to deliver future trainings for new or novice teacher supervisors.

Closure

McREL recommends that placing a system to support and evaluate teachers is a central feature of a talent identification, development, and support strategy. McREL recommends consideration of a standardized and consistent method to support, evaluate, and retain quality teachers. Another recommendation is to re-examine the purpose of evaluation in order to enhance the utility of the system. Furthermore, emphasize negotiable expectations for the consistent application of the system requiring fidelity to the processes, procedures, and protocols. Improving teacher evaluation is not an end in itself, but should be the cornerstone that communicates and expects sound teacher practice and desirable educational outcomes for students.

REFERENCES

Beesley, A. D., & Apthorp, H. S. (2010). Classroom Instruction That Works: Research Report. *Mid-continent Research for Education and Learning (McREL)*.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. ASCD.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education policy analysis archives*, *8*, 1.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, *38*(2), 109-119.

Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.