SRI International

WestEd

Examining Classroom Observation Rubric Data

Issues emerging from classroom observation rubric data submitted August 2017

Published in collaboration with SRI International

WestEd is a research, development, and service agency whose mission is to promote excellence, achieve equity, and improve learning for children, youth, and adults. For more information about WestEd, visit http://www.wested.org/; call 415.565.3000 or, toll-free, (877)4-WestEd; or write: WestEd / 730 Harrison Street / San Francisco, CA 94107–1242.

SRI Education, a division of SRI International, is tackling the most complex issues in education to identify trends, understand outcomes, and guide policy and practice.

SRI International is a nonprofit research institute whose innovations have created new industries, extraordinary marketplace value, and lasting benefits to society. SRI International is a registered trademark and SRI Education is a trademark of SRI International. All other trademarks are the property of their respective owners.

Suggested citation: Torre, D., Gallager, A., White, M., (2017). *Examining classroom observation rubric data: Issues emerging from classroom observation rubric data submitted in August 2017.* San Francisco, CA: SRI International and WestEd.

Contents

Introduction

The New Generation of Educators Initiative (NGEI), funded by the S.D. Bechtel, Jr. Foundation (hereafter "the Foundation"), seeks to strengthen the current teacher preparation system in California so that new teachers enter the workforce prepared to implement Common Core State Standards (CCSS) and the Next Generation Science Standards (NGSS). The Foundation has developed a theory of action to guide reform that focuses on five Key Transformation Elements (KTEs): partnership (KTE 1), prioritized skills (KTE 2), practice-based clinical preparation (KTE 3), formative feedback on prioritized skills (KTE 4), and data-driven continuous improvement (KTE 5). WestEd and SRI International are conducting a formative evaluation to track NGEI implementation and outcomes at the CSU campuses that have received grants, in both Phase 1 (March 2015 – June 2016) and in Phase 2 (June 2016 – June 2019) of the initiative.

Campuses participating in Phase 2 of the NGEI initiative were required to adopt or create a classroom observation rubric to measure prioritized skills jointly identified and agreed upon by campuses and their partner districts, and use the rubric to assess candidates' development of prioritized skills on a regular basis during clinical placements.

The data generated from the classroom observations are intended to serve multiple purposes:

- To anchor feedback candidates receive from faculty, university supervisors, and cooperating teachers to a common set of indicators to ensure consistent focus on high quality teaching practices.
- To assess candidate performance overall and on specific dimensions for the purpose of identifying areas for targeted support.
- To support program-level improvements by identifying trends in candidates' strengths and weaknesses.

In order to serve any of these purposes, observational rubrics should be designed to measure candidate progress toward prioritized skills in a valid and reliable way. Importantly, the observational rubric must accurately capture variation between the dimensions of teaching included on the rubric, among candidates, and over time.

In August 2017, the Foundation asked NGEI campuses to submit observation rubric data from all or a subset of candidates enrolled in funded programs from the most recent semester from which data was available and to write a brief reflection on their rubric data. Campuses could choose to submit data from one or more points in time. The Foundation requested that WestEd/SRI analyze the data submitted by NGEI sites.

The purpose of this memo is to provide an overview of the data campuses submitted, highlight patterns in the data, and identify issues that can inform changes in how campuses use their rubrics to support candidates.

Background

NGEI's requirement that all NGEI partnerships select classroom observation rubrics for use in teacher preparation programs by September 1, 2016 represented a major first activity for the NGEI partnerships that had received grants just three months earlier. After adoption of the classroom observation rubrics, NGEI partnerships were expected to begin implementation of the rubrics such that candidates would receive consistent feedback on progress toward prioritized skills.

In 2016–17, NGEI campuses made strides in the selection/development and implementation of classroom observation rubrics. As documented below in Exhibit 1, all but two campuses had their new observational rubrics in place in NGEI Year 1, 2016–17; eight NGEI campuses launched training efforts to prepare university supervisors, cooperating teachers, and/or faculty to use the rubrics to observe candidates and provide feedback; and four NGEI campuses used the rubrics as part of their regular feedback process.

Exhibit 1. Classroom Observation Rubrics Use Overview

Campus	Rubric	Implementation Activities as of Spring 2016–17
Bakersfield	Danielson-based	Supervisors and cooperating teachers participated in regular training and began observing candidates in fall 2016.
Cal Poly, SLO	Danielson-based	Supervisors participated in two trainings focused on norming and began observing candidates in fall 2016.
Channel Islands	Locally developed	Channel Islands did not finalize its rubric or provide related training. Its tentative plan was to begin training and pilot using the rubric in fall 2017.
Chico	TNTP	Supervisors piloted using rubrics with candidates but did not use observations to provide feedback to candidates. Supervisors and faculty participated in online calibration training.
Dominguez Hills	Locally developed	Dominguez Hills did not finalize its rubric and did not provide related training. It planned to provide training for cooperating teachers in spring and summer 2017 and begin observing candidates using the rubric in fall 2017.

Campus	Rubric	Implementation Activities as of Spring 2016–17
Fresno	Locally developed	Fresno piloted its rubric in 2016–17 but did not provide related training. It planned to provide 10 hours of training for cooperating teachers and supervisors in summer 2017 and begin observing candidates using the rubric in fall 2017.
Fullerton	MCOP 2	Fullerton piloted the MCOP2 in 2016–17 and provided feedback using observations to one candidate. Supervisors, clinical coaches, faculty, and candidates participated in six rubric-related trainings.
Long Beach	Locally developed	Supervisors and cooperating teachers participated in one rubric training session and began observing candidates in spring 2017.
Monterey Bay	Locally developed	Candidates were not formally observed using the rubric, but supervisors were encouraged to provide feedback using language from the rubric, and one faculty member used it in her methods class. Supervisors participated in ongoing rubric training.
Sacramento	Locally developed	Members of the instructional leadership team participated in two rubric training sessions. Sacramento planned to begin observing candidates using the rubric in fall 2017.
Stanislaus	5D+	Supervisors participated in two trainings focused on norming and began observing candidates in spring 2017.

By the time NGEI campuses submitted data to the Foundation in August 2017, they had been using their rubrics to observe candidates for at most two semesters. Several campuses (e.g. Monterey Bay, Chico, Fullerton) considered 2016–17 to be a pilot year and did not intend for observations to be used for evaluative, or even formative, feedback purposes. Exhibit 2 shows when each campus began using their classroom observation rubric for feedback or pilot purposes.

Exhibit 2. Timeline of Implementation of Classroom Observation Rubrics

Fall 2016

Bakersfield
Cal Poly, SLO
Monterey Bay*
Chico*

Spring 2017

Long Beach
Stanislaus
Fullerton*

Fall 2017**

Channel Islands
Dominguez Hills
Sacramento
Fresno

*Note: Chico, Fullerton, and Monterey Bay used their rubric for pilot purposes but did not typically share the observation ratings with candidates.
**Note: During our spring 2017 site visits, Channel Islands, Dominguez Hills, Sacramento, and Fresno shared their plans to begin observing candidates using the classroom observation rubrics in Fall 2017.

Given that implementation was in its initial phases in the 2016–17 year, the Foundation expected there to be measurement and implementation issues that would be evident in the data and asked WestEd/SRI to conduct analysis of early data to inform Foundation support and implementation at NGEI sites in the coming months. We report trends emerging from the early implementation phase below.

Data Submitted to the Foundation

Seven of 11 NGEI campuses submitted classroom observation rubric data gathered during the 2016–17 school year. All campuses that submitted ratings (hereafter, campuses) used rubrics that had a functional range from 1–4.[1] (See Appendix A for sample information for each campus.) Rubrics at four campuses were specifically developed or adapted to assess the performance of pre-service teachers, while the remaining campuses adopted rubrics designed to assess the performance of in-service teachers. Campus rubrics included from four to 34 indicators (i.e., individual dimensions of effective teaching). Campuses submitted ratings for 13 to 64 candidates and each candidate was observed one to five

---

[1] The rubric used at Campus C uses a five-point scale; however, Campus C has decided not to give ratings of 5 to candidates because they are novice teachers. Campus A allows for raters to give half points on their rubric, which creates a 7-point scale; however, the ratings range from 1–4, making them comparable to other campuses. Campus B's rubric uses a 0 to 3 scale, which means it also has 4 points. For the purposes of this report, all of their ratings have been shifted to a scale of 1 to 4.

times (at two campuses, some candidates were observed more times than others) for a total of 26 to 94 observations represented by the data. The number of ratings on individual indicators that campuses submitted ranged from 108 to 3,196.[2] In the remainder of this report as we discuss trends in the observation rubric data, we have anonymized campus names.

Themes Emerging from Early Implementation Data

Overall, we find that while the ratings on most campuses suggest differences in candidate skills, ratings appear inflated on most campuses. This is of concern because inaccurate measures of candidate skills could hinder campuses' ability to identify candidates in need of additional support, reduce the efficacy of feedback that would help candidates improve, and dim the signals campuses receive about areas of strength as well as those in need of improvement in their own programs. Many NGEI project leads are already aware of this issue and are working to improve it by providing additional training for observers, revising processes for observing candidates and collecting ratings data, or making modifications to the rubric.

At most campuses, observation ratings were skewed towards the highest levels of performance.

Research using observational rubrics where highly calibrated evaluators assessed the performance of large samples of teachers shows that the distribution of teacher performance follows a bell curve, with most teachers receiving ratings in the middle of the distribution.[3,4] In contrast, **at nearly all campuses, the distribution of ratings was skewed positively** as shown in Exhibit 3 below. Notably, at all but one campus, more than two-thirds of ratings were in the top two categories of performance (3 or 4).

---

[2] Campus A submitted the average rating provided by cooperating teachers and/or university supervisors during the spring 2017 semester for each candidate. That is, each of the 3,196 ratings from Campus G represents an average across multiple ratings. Campus A did not provide an indication of the total number of individual ratings.

[3] Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

[4] The research cited above was conducted using a sample of professional teachers, not teacher candidates. We expect that on average, candidates would receive lower ratings than professional teachers. Of note, Campus B, Campus C, Campus D, and Campus F use rubrics designed to measure the instructional quality of professional teachers while Campus A, Campus E, and Campus G use rubrics designed to measure the instructional quality of teacher candidates. Despite the differences in the intended population, the modal rating for all but one campus (Campus G) is a three or four.

Thirty-four percent or fewer ratings were level 2, and less than 10 percent of ratings were at the lowest level (1).

Exhibit 3. Distribution of Ratings

Percent of Ratings
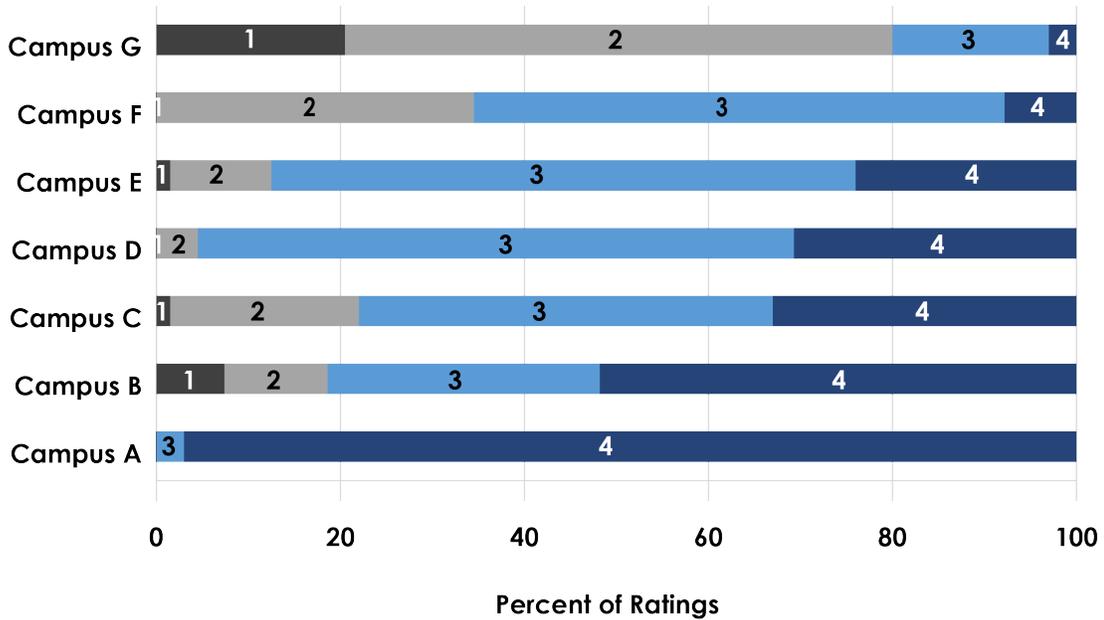
Note: Each rating submitted by Campus A represented an average of multiple ratings given by cooperating teachers and/or university supervisors during the spring 2017 semester. While it is likely that the distribution of all ratings would look slightly different than what is presented in Exhibit 3, the general positive skew would persist.

The distribution of ratings was compressed to a greater degree at some campuses than others.[5] At Campus G, the modal rating was two, about 38 percent of candidates received a 1 or a 3, and only 3 percent of candidates received the highest rating (n=108 ratings), which follows a relatively expected distribution of ratings of teacher candidates. In contrast, at Campus A, 91 percent (2,894 of a total of 3,196) of ratings were between 3.5 and 4 (indicated by a 4 in Exhibit 3; 3 indicates ratings between 3.0 and 3.49),[6] and fewer than 1 percent of candidates received any rating lower than a 3 (25 out of a total of 3,196). If most candidates are receiving the same ratings, as is the case at Campus A and, to a lesser extent, other campuses, the data are likely not capturing differences in candidate performance.

---

[5] We would expect candidates to receive higher ratings as they gain more experience; therefore, we would expect ratings of performance in the spring to be higher than ratings of performance in the fall. Campuses did not consistently indicate when observations were conducted and so for the purposes of this memo we did not analyze how the timing of observations was related to the distribution of ratings.

[6] Ratings at Campus A were not always even numbers (1–4) because 1) observers could give half-point ratings and 2) ratings represent an average of multiple ratings.

Inflated ratings for individual candidates masked potential variation in their performance across rubric indicators. To illustrate, Exhibit 4 shows the ratings for a Campus B candidate who was observed three times. These data show how high initial ratings created a *ceiling effect*; that is, because the candidate received the highest rating (4) on five of nine indicators for their first observation, there was little room for the candidate to show improvement during subsequent observations. While this was the only Campus B candidate who was observed three times, six of the 23 candidates observed at Campus B had one observation where they received a 4 on every indicator, and similar instances occurred across all campuses.

**Exhibit 4. Ratings by Indicator for One Candidate, Campus B**

Observation	Candidate 1		
	1	2	3
1. Students engage in exploration/investigation/problem solving.	4	4	4
2. Students used a variety of means to represent concepts.	4	4	4
3. Students were engaged in mathematical activities.	4	4	4
4. Students critically assessed mathematical strategies.	2	4	4
5. Students persevered in problem solving.	4	4	4
6. A high proportion of students talking related to mathematics.	2	4	4
7. The teacher created a climate of respect for what others had to say.	2	4	4
8. The teacher provided wait-time (think-time).	4	4	4
9. Students were involved in the communication of their ideas.	2	4	4

Inflated ratings risk communicating to candidates that they do not have areas for improvement. The data shown in Exhibit 4 suggest that this candidate mastered the majority of indicators in their first observation and had mastered all of the prioritized skills measured by the rubric by the second observation. The ratings provide little indication to the candidate about how to improve, nor to faculty, cooperating teachers, or supervisors about how they can support this candidate.

Inflated ratings for individual candidates translated into limited variation in average performance across indicators, potentially obscuring where campuses should focus program improvement efforts. Ideally, variation in ratings for each

indicator aggregated across candidates could help program leaders identify areas for program improvement by highlighting areas of strength and weakness. As an example of the typical pattern found at NGEI campuses, Exhibit 5, below, shows the average rating aggregated across candidates for each indicator at Campus E. The data show that the highest rated indicators across candidates were related to classroom management, relationships, and professionalism, while the lowest rated indicators include supporting emergent bilinguals (EBs), setting instructional outcomes, and questioning and discussion techniques. These data follow a pattern found in other studies whereby teachers tend to perform highest on the least complex teaching tasks (e.g., classroom management, developing relationships) and lowest on more complex or abstract dimensions of teaching, such as promoting critical thinking, questioning, and differentiating instruction.[7] The data could equip program leaders at Campus E to reflect upon what drives the variation in performance across different indicators; for instance, why candidates are rated lowest on indicators related to supporting emergent bilinguals or what the program is doing well to cultivate candidates' ability to be reflective, respectful practitioners.

Exhibit 5. Average Rating Across Indicators, Campus E
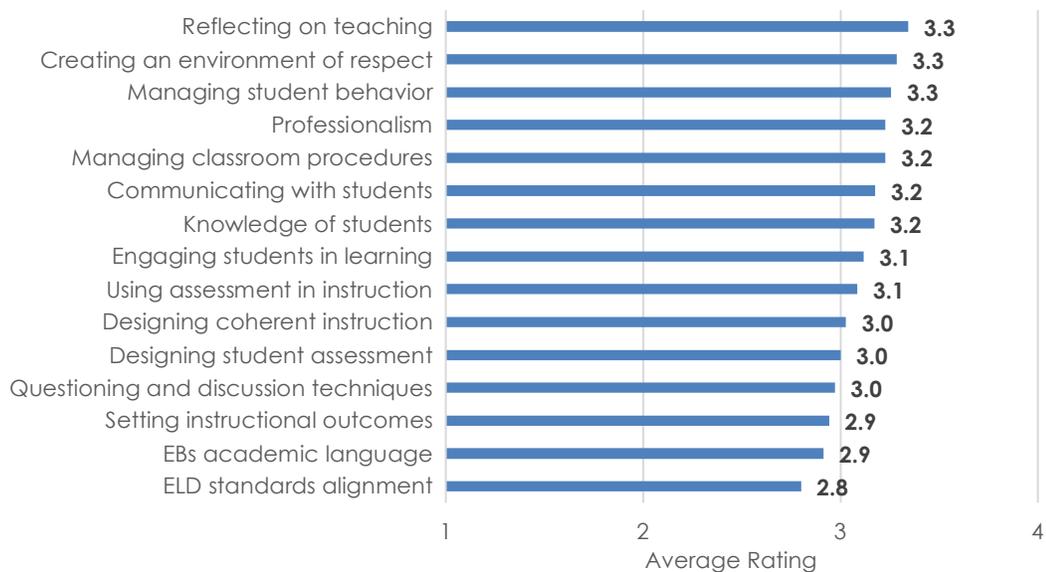
Despite some evidence to suggest areas of program strength and weakness, the data in Exhibit 5 offer minimal evidence that there are meaningful differences in candidates' mastery of different indicators. In the case of Campus E, the difference between the

---

[7] Kane & Staiger, 2012.

highest and lowest rated indicator across candidates was only .5, and the average rating for 10 of the 13 indicators was at least a 3, indicating that candidates *demonstrated proficiency*.

The skewed distribution of ratings may be a result of rater error or lack of calibration.

Positively skewed ratings might indicate that observers were making any number of rating errors or were not calibrated around a set of high expectations for candidate performance. Issues impacting observers' ability to provide accurate ratings include:

- **Familiarity bias.** A bias towards higher or lower ratings of a particular teacher based on the rater's personal relationships with that person.[8,9] For example, university supervisors at one campus reported that they felt uncomfortable giving low scores (e.g., 1s and 2s) and so tended to inflate ratings.

- **Rater drift.** A rating error whereby observers rate teachers higher over time, regardless of actual performance. Rater drift can happen if observers have not had regular calibration training to ensure that their ratings continue to be valid over time and in different contexts.

- **Unclear expectations.** At most campuses, observer training has been focused on norming (i.e., developing a shared understanding of performance levels) rather than calibration (i.e., ensuring observers rate performance within a particular threshold). The process of norming is a necessary first step in training observers to use a rubric, but can allow for groupthink to muddle expectations. For example, in the process of developing a shared understanding, more critical observers may lower their expectations to adhere to the group consensus.

A skewed distribution limits how rubric data can be used to inform candidate feedback and program improvement.

The evidence suggests that in many cases, ratings were not valid or reliable measures of candidate skills, which can have negative consequences for how the data can be used, including:

---

[8] Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings.* Washington, DC: Center for Educator Compensation Reform. Retrieved from https://files.eric.ed.gov/fulltext/ED532068.pdf.

[9] Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study.* Seattle, WA: Bill & Melinda Gates Foundation.

- Supervisors may not have sufficient data to provide candidates with the consistent and targeted feedback they need to progress toward mastery of prioritized skills.
- Candidates and campuses may not be able to accurately assess candidates' progress over time.
- Candidates may be graduating with an unrealistic sense of their mastery of prioritized skills and the continued improvement that will be necessary to achieve teaching excellence.
- There may not be enough variation in the data to identify the relative strengths and weaknesses of the program.

Conclusion

In the 2016–2017 academic year, NGEI campuses made considerable progress toward creating systems for assessing candidate progress toward prioritized skills using classroom observation rubrics, and are engaged in continuously assessing and improving these systems. The findings in this brief, based on data from the earliest stages of implementation, are intended to promote reflection and to inform efforts to improve the value of the rubrics in support of candidate development and continuous program improvement.

Early implementation data show that ratings were consistently positively skewed, a trend that can be problematic if the skewed ratings discourage honest reflection from all stakeholders, including candidates, supervisors, cooperating teachers, and faculty, around what constitutes good teaching and mastery of prioritized skills. To address the potential threats to using rubric data to inform feedback to candidates or program improvement that arise from using skewed data, campuses can:

- **Clarify theories about how classroom observation rubric data can be used to support candidates in learning prioritized skills.** The rubric data should be used as part of an intentional and ongoing process[10] of introducing prioritized skills, providing opportunities to practice those skills, and providing feedback on progress toward those skills. To strengthen that process, campuses might map out the connection between the adopted rubric and programs' prioritized skills; when and how feedback will be shared with candidates; and how candidates might be supported to act on the feedback to further develop prioritized skills.
- **Recalibrate expectations for novice teacher performance.** The stated goal at several campuses is for a large majority of candidates to receive ratings of 3 or 4 on observation rubrics with a 4-point scale by the end of their program.

---

[10] See Exhibit B-1 in the Appendix for an illustration of this process.

At campuses that designed their rubrics specifically for use with novice, pre-service teachers,[11] this might be a realistic goal. However, most campuses selected a rubric that was designed for in-service teachers, and so a more realistic expectation might be that candidates average between 2 and 3 on a 4-point scale by the end of their first year.

- **Provide ongoing training to ensure that ratings are valid and reliable.** As campuses continue implementing observer training, they should ensure that observers are not only calibrated with one another, but also calibrated to a true rating for each level of performance on their rubric. It is important that calibration training at all campuses, even those who intentionally use their observation rubric to provide non-evaluative, formative feedback, be ongoing and designed to continuously mitigate the threats of rater drift, familiarity bias, or other sources of rater error.

Many campuses have acknowledged the issues outlined in this report and are actively working to address them. To ensure that all campuses are using their rubrics to inform the development of candidates and improve their program, we recommend a focus on setting realistic expectations and calibrating observers around those expectations in the coming months.

[11] See Exhibit A-1 in the Appendix for a list.

Appendix A: Notes on Rubrics and Sample

Exhibit A-1 provides sample information for each campus. Rubrics at CSU Long Beach, CSU Monterey Bay, and Cal Poly, SLO were specifically developed or adapted to assess the performance of pre-service teachers. The remaining campuses adopted rubrics designed to assess the performance of in-service teachers. At Bakersfield and Chico, some candidates were observed more often than others. At Long Beach, candidates were observed at least twice by either a cooperating teacher or university supervisor.

Exhibit A-1. Rubric and Sample Details

Campus	Rubric	Indicators on rubric	Candidates observed	Observations per candidate	Observations submitted [c]	Ratings submitted [d]
Bakersfield	Danielson-based	10 [b]	16	1–5	33	123
Chico	TNTP	4	16	1–2	31	124
Fullerton	MCOP 2	9	24		31	277
Long Beach	Clinical Practice Evaluation Form [a]	34	64	2	94	3,196
Monterey Bay	STEM Teaching Rubric [a]	3	20	2	36	108
SLO	SOE Tool [a]	15	18	3	35	523
Stanislaus	5D+	30	13	2	26	711

[a] Rubric was designed to assess the performance of pre-service teachers.
[b] Bakersfield submitted data for 5 of the 10 indicators on their rubric as part of their August 2017 report.
[c] Observations submitted shows the total number of observations across all candidates (classroom observation events * number of candidates) that were submitted at each campus.
[d] Ratings submitted shows the total number of individual ratings (indicators * classroom observation events * number of candidates) that were submitted at each campus.

Exhibit A-2 shows the labels for each level of proficiency for each campus' rubric. Chico does not use the 5th point on the scale. Fullerton's scale ranges from 0–3; the scale has been shifted for the purposes of this report.

Exhibit A-2. Proficiency Labels

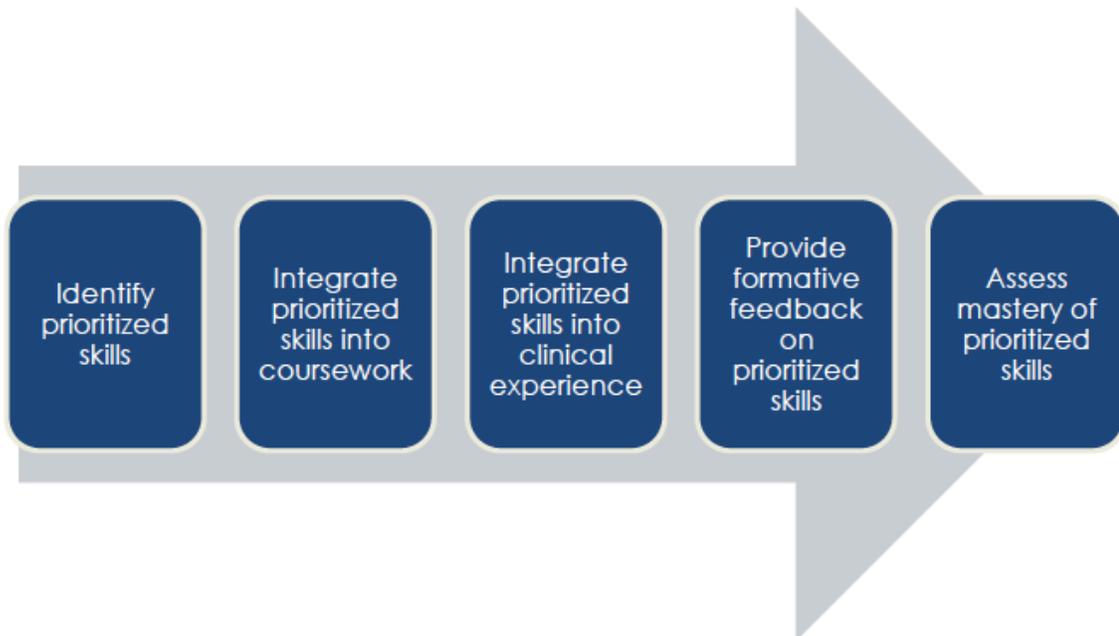| Campus	1	2	3	4	5
Bakersfield	Unsatisfactory	Basic	Proficient	Distinguished	n/a
Chico	Ineffective	Minimally effective	Developing	Proficient	Skillful
Fullerton	Ineffective	Developing	Effective	Highly effective	n/a
Long Beach	Not proficient	Attempting	Developing	Proficient	n/a
Monterey Bay	Ineffective	Emerging	Practicing	Applying	n/a
SLO	Did not demonstrate	Partially demonstrated	Demonstrated	Demonstrated with distinction	n/a
Stanislaus	Unsatisfactory	Basic	Proficient	Distinguished	n/a

Appendix B: NGEI Design Elements

Exhibit B-1 shows how the KTEs work together to ground a process by which prioritized skills are selected, introduced, practiced, and assessed.

Exhibit B-1. NGEI Design Principles for Identifying and Integrating Prioritized Skills into Quality Teacher Preparation Programs

Purpose

This document introduces a set of design principles for guiding the development of a set of skills that teacher candidates will know and be able to enact by the end of their teacher preparation program. The principles are grounded in the practical expertise of NCTR and in research-based evidence about the key characteristics of effective professional learning programs. The principles represent a manageable, versatile set of design criteria that can inform the assessment, development, and improvement of quality teacher residencies. To this end, they can also serve as a basis for benchmarking and learning across diverse programs.

Overview of the Design Principles for Establishing Prioritized Skills

Identify prioritized skills → Integrate prioritized skills into coursework → Integrate prioritized skills into clinical experience → Provide formative feedback on prioritized skills → Assess mastery of prioritized skills

Explanation of Design Principles for Establishing Prioritized Skills

Identify prioritized skills	Integrate prioritized skills into coursework	Integrate prioritized skills into clinical experience	Provide formative feedback on prioritized skills	Assess mastery of prioritized skills
Programs should have 7–10 prioritized skills. Prioritized skills (e.g. leading a group discussion, eliciting and interpreting student ideas, building relationships with students) are teaching skills of moderate grain size that programs have selected as the most important for candidates to learn during their preparation program. Skills are observable and NGEI encourages campuses to measure candidate progress towards prioritized skills.	An authentic integration of prioritized skills into coursework. This includes, but is not limited to, how teacher candidates learn prioritized skills through the lesson/unit planning process, enactment of rehearsals in coursework, and studies of enactment (Example: Teaching and Learning Cycle)	Candidates are placed in training site schools with quality mentor teachers who model the prioritized skills and can deliver quality feedback based upon them.	After they have been introduced to the skills and have had multiple opportunities to practice them in courses and then in placements, candidates are observed with classroom observation rubrics. Using data from the observations and other data on candidate progress, supervisors and mentor teachers give aligned, specific evidence-based and frequent feedback to candidates to build upon prioritized skills.	Programmatic assessments (or multiple measures) put into place to determine whether teacher candidates are able to enact the program's identified prioritized skills to inform any needed remediation and determination of completion.
	Candidate learning of prioritized skills based on the learning cycle	Candidate learning of prioritized skills based on the learning cycle	Candidate learning of prioritized skills based on the learning cycle	