NGEI Issue Brief

SRI International

WestEd

Examining Classroom Observation Rubric Data

Issues emerging from classroom observation rubric data submitted in August 2017

Daniela Torre
Alix Gallagher
Melissa White

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.

This publication was made possible by a grant from the S. D. Bechtel, Jr. Foundation via its "Preparing a New Generation of Educators for California" initiative. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the Foundation.

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.

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 received comprehensive grants in Phase 2.

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 create a common understanding of high quality teaching practices that can serve to anchor feedback candidates receive from faculty, university supervisors, and cooperating teachers
- 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 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 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 the future collection and use of classroom observation data to strengthen supports to candidates.

Data

Seven of 11 NGEI campuses submitted classroom observation rubric data gathered during the 2016-17 school year. All campuses that submitted ratings data (hereafter, campuses) used rubrics that had a functional range from 1–4.[1] (See appendix for sample information for each campus). Campus rubrics included from four to 34 indicators (i.e. individual dimensions of effective teaching). Campuses submitted data for 13 to 64 candidates and each candidate was observed one to five times (at Bakersfield and Chico, 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]

Findings

Overall, we find that while the ratings on most campuses suggest differences in candidate skills, ratings appear inflated on most campuses. 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.

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] In contrast, **at nearly all campuses, the distribution of ratings was skewed positively**, as shown in Exhibit 1 below. Notably, at all but one campus, more than two-thirds of ratings were in the top two categories of performance (3 or 4). Thirty-

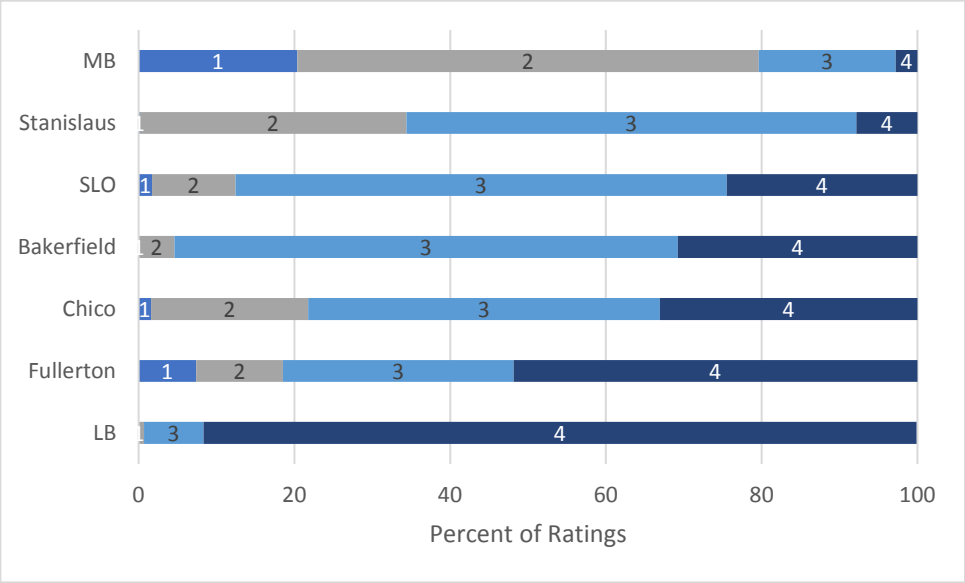
[1] The TNTP rubric, used at CSU Chico, uses a five-point scale; however, Chico has decided not to give ratings of 5 to candidates because they are novice teachers. CSU Long Beach allows for raters to give half points on their rubric, which creates a 7-point scale; however, the ratings range from 1-4, making them comparable to other campuses. Fullerton'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.

[2] CSU Long Beach submitted the average rating provided by mentor teachers and/or university supervisors during the spring 2017 semester for each candidate. That is, each of the 3,196 ratings from Long Beach represents an average across multiple ratings. Long Beach 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.

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

Exhibit 1. Distribution of Ratings



Note: Each rating submitted by CSU Long Beach represented an average of multiple ratings given by mentor 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 1, the general positive skew would persist.

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

The positively skewed distribution of ratings was also reflected in inflated ratings for individual candidates, which masked potential variation in performance across

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

rubric indicators. To illustrate, Exhibit 2 shows the ratings for a CSU Fullerton 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 her first observation, there was little room for her to show improvement during subsequent observations. While this was the only Fullerton candidate who was observed three times, 6 of the 23 candidates observed at Fullerton had one observation where they received a 4 on every indicator and similar instances occurred across all campuses.

Exhibit 2. Ratings by indicator for one candidates, Fullerton

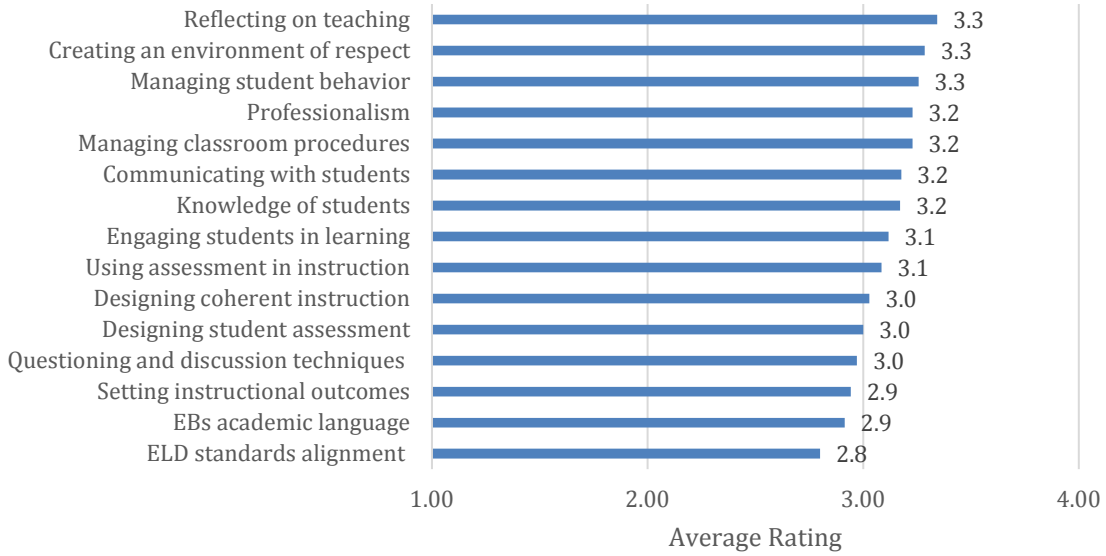
		Candidate 1		
	Observation	1	2	3
1.	Students engage in exploration/investigation/ problem solving.	4	4	4
2.	Students used a variety of means to represent concepts.	4	4	4
3.	Students were engaged in mathematical activities.	4	4	4
4.	Students critically assessed mathematical strategies.	2	4	4
5.	Students persevered in problem solving.	4	4	4
6.	A high proportion of students talking related to mathematics.	2	4	4
7.	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

Finally, inflated ratings risk communicating to candidates that they do not have areas for improvement. The data shown in Exhibit 2 suggest that this candidate mastered the majority of indicators in her first observation and had mastered all of the prioritized skills measured by the rubric by her second observation. The data provides little indication to the candidate about where she can improve or to faculty, mentor teachers, or supervisors about how they can support this candidate.

Variation in ratings for each indicator aggregated across candidates can be useful for identifying areas for program improvement. As an example of the typical pattern found at NGEI campuses, Exhibit 3, below, shows the average rating for each indicator at Cal Poly, San Luis Obispo (SLO). 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. **This data follows a pattern found in other studies whereby teachers tend to perform highest on the least complex teaching tasks (e.g. classroom management, developing relationships, etc.) and lowest on more complex or abstract dimensions of teaching, such as**

promoting critical thinking, questioning, and differentiating instruction.[5] The data can equip program leaders at Cal Poly, SLO 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 3. Average rating across indicator, Cal Poly, SLO



Despite some evidence to suggest areas of strength and weakness for individual candidates as well as for programs, the generally inflated ratings translated into limited variation in average performance across indicators. In the case of Cal Poly, SLO, the difference between the 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*. These data could be interpreted as showing that there are actually few areas for program improvement as the program already is adequately preparing candidates, when, in reality, there are likely many areas where the campus could work to improve how they support candidate progress towards mastery of prioritized skills.

[5] Kane & Staiger, 2012

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.[6,7] 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 "group think" 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 that:

- Supervisors may not have sufficient data to provide candidates with the consistent and targeted feedback they need to progress towards mastery of prioritized skills.
- Candidates and campuses may not be able to accurately assess candidates' progress over time.

---

[6] 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 http://cecr.ed.gov/pdfs/ Inter_Rater.pdf

[7] 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.

- 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 their program.

Conclusion

Consistently positively skewed ratings discourage honest reflection from all stakeholders, including candidates, supervisors, mentor teachers, and faculty, around what constitutes good teaching and mastery of prioritized skills. If all pre-service candidates are consistently earning high ratings, it suggests that the bar for excellent teaching is too low and clouds potential areas for improvement. To address these various threats to using rubric data to inform feedback to candidates or program improvement, campuses can:

- Clarify theories about the role of classroom observation rubric data in informing supports to ensure candidates learn prioritized skills. The rubric data should be used as part of an intentional and ongoing process[8] of introducing prioritized skills, providing opportunities to practice those skills, and providing feedback on progress towards those skills. To strengthen that process, campuses might map out when particular skills will be assessed in relation to when they are introduced and practiced; 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.
- Re-calibrate 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 the observation rubric by the end of their program. At campuses who designed their rubrics specifically for use with novice, pre-service teachers,[9] 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 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

[8] See Exhibit A-3 in the Appendix for an illustration of this process.

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

campuses, even those who intentionally use their observation rubric to provide non-evaluative, formative feedback, be ongoing and designed to continuously mitigate against the threats of rater drift, familiarity bias, or other sources of rater error.

NGEI campuses have made great strides in terms of creating systems for assessing candidate progress towards prioritized skills. Further, many campuses have acknowledged the issues outlined in this report and are actively working to address them. To ensure that all campuses are collecting data that are useful for informing the development of candidates and program improvement, campuses should work towards setting realistic expectations and calibrating observers around those expectations in the next year of the initiative.

Appendix A: Notes on Rubrics and Sample

Exhibit A-1 provides sample information for each campus. Rubrics at CSU Long Beach, CSU Monterrey 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 mentor teacher or university supervisor. The last column (Ratings submitted) shows the total number of individual ratings (indicators * classroom observation events * number of candidates) that were submitted at each campus.

Exhibit A-1. Rubric and Sample Details

Campus	Rubric	Indicators on rubric	Candidates observed	Observations per candidate	Observations submitted	Ratings submitted
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.

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

	Bakersfield	Chico	Fullerton	Long Beach	Monterrey Bay	SLO	Stanislaus
1	Unsatisfactory	Ineffective	Ineffective	Not proficient	Ineffective	Did not demonstrate	Unsatisfactory
2	Basic	Minimally Effective	Developing	Attempting	Emerging	Partially Demonstrated	Basic
3	Proficient	Developing	Effective	Developing	Practicing	Demonstrated	Proficient
4	Distinguished	Proficient	Highly Effective	Proficient	Applying	Demonstrated with Distinction	Distinguished
5		Skillful					

Appendix B: NGEI Design Elements

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

Exhibit B-1. NGEI Design Elements – In development -

Identify prioritized skills	Integrate prioritized skills into coursework	___	__	__
A small set of 5-10 essential specific prioritized and observable skills that are agreed upon by the program and district partners.	In development		T	