BRIEF 4: EXAMINING THE USE OF SLOS TO EVALUATE TEACHERS

CADRE
Center for Assessment Design
Research and Evaluation (CADRE)
University of Colorado, Boulder

Key highlights from brief:

- SLOs based on English Language Arts (ELA) and Math district templates were among the most popular with teachers, and SLOs with an ELA focus were significantly more prevalent than those with a Math focus.
- Most students tracked on SLOs received 2 or 3 growth points out of a maximum of 3 points. This suggests that, according to their teachers, most students showed evidence of moderate to high growth.
- Most teachers earned effective or distinguished ratings using the LEAP rules for calculating teacher growth.
- There is evidence to suggest that teachers were underestimating the preparedness classifications of some of their students, which would tend to inflate both student and teacher-level growth scores.
- Teacher-level measures of growth based on SLO scores were not associated with classroom demographic characteristics and prior year student performance on the PARCC tests. It does not appear to be the case that it was easier or harder to demonstrate growth with certain kinds of classrooms.
- Teacher-level measures of growth were significantly associated with teacher professional practice ratings. This suggests, tentatively, that teacher-level SLO measures may be capturing a characteristic of teachers related to their ability as instructors.

OVERVIEW

As discussed in Brief 2, a challenge in having teachers integrate SLOs as a critical part of their regular classroom practices is the fact that SLOs are intended to serve dual purposes. On the one hand, there is a desire for SLOs to be used formatively to inform instructional practice. On the other hand, SLOs provide teacher-level measures that can count both toward a teacher's monetary compensation (i.e., ProComp) and their annual evaluation (i.e., LEAP). In this brief we focus on the use of SLOs for this latter purpose. There are three important threats to the validity of SLOs as teacher-level measures of student growth.

1. First, unlike student growth percentiles (SGPs) which are based on objective and standardized measures of student achievement, SLOs are based upon more subjective measures of student achievement in that they are mediated by teacher scoring decisions. This could create an incentive for teachers to inflate student growth either by classifying some students as less prepared for an SLO at the outset of an instructional period than they actually are, or by classifying students at a higher level of mastery than they have actually attained.
2. Second, even if teachers are not intentionally inflating student growth scores, it may be the case that it is easier to demonstrate growth with certain kinds of students in certain kinds of classroom contexts. For example, some teachers might be concerned that it would be harder to show growth with a classroom comprised of a majority of gifted and talented students relative to a classroom comprised of a majority of students with IEPs. If this were to be the case, then teacher-level SLOs would be biased by classroom contexts that are outside of a teacher's control.
3. Third, even if SLOs are not biased, it may be the case that they are subject to so much measurement error that they will be volatile from year to year, such that the same teacher might shift from a high score one year to a low score the next year just because they had a strong student cohort one year and a weak one the next.

In this brief we examine evidence relevant to the first two of these three threats to validity. It is not possible to evaluate the third threat to validity until we have two years of SLO data for DPS teachers. We use information about prior year student performance on Colorado's state-administered assessment (PARCC tests in Math and English Language Arts (ELA)) as a criterion against which the SLO preparedness of students can be compared. Specifically, we examine whether students who scored in the "met expectations" or "exceeds expectations" performance levels on a PARCC test in math or ELA for their previous grade are classified as "prepared" or "ahead" for their same subject SLO in the following grade. We consider a mismatch between a PARCC classification and a preparedness classification as evidence of an inconsistent classification. We then examine whether student characteristics are associated with the probability of an inconsistent preparedness classification.

To evaluate whether teacher-level SLO scores are biased against teachers in certain classroom contexts, we examine whether these scores vary as a function of classroom characteristics such as prior achievement, percent of nonwhite students, free and reduced lunch status, etc. We also contrast this to an analysis of a teacher-level variable that is an aggregate of student SLO mastery without factoring in differences in student preparedness. Finally, we examine the relationships of these variables with scores on LEAP professional practice observation protocols.

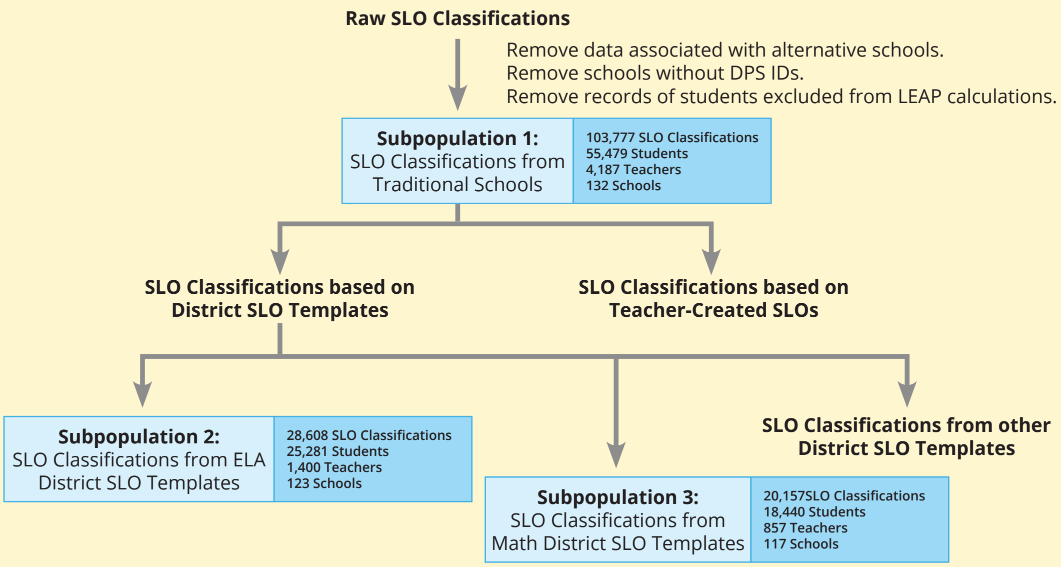
DATA

Our sample consisted of all district schools that implemented the SLO process in 2015-16. Alternative schools were excluded. Also excluded were individual students who were not assigned SLO growth ratings due to factors such as attendance. In the analyses that follow, we distinguish between district-created SLOs for ELA and Math.[1] Of interest are three subpopulations:

1. Students with SLOs from Traditional Schools
2. Students with ELA SLOs from District Templates
3. Students with Math SLOs from District Templates

A flowchart depicting these three subpopulations is summarized in Figure 1.

FIGURE 1. SLO SUBPOPULATIONS


Raw SLO Classifications
Remove data associated with alternative schools.
Remove schools without DPS IDs.
Remove records of students excluded from LEAP calculations.
Subpopulation 1:
SLO Classifications from Traditional Schools
103,777 SLO Classifications
55,479 Students
4,187 Teachers
132 Schools
SLO Classifications based on District SLO Templates
SLO Classifications based on Teacher-Created SLOs
Subpopulation 2:
SLO Classifications from ELA District SLO Templates
28,608 SLO Classifications
25,281 Students
1,400 Teachers
123 Schools
Subpopulation 3:
SLO Classifications from Math District SLO Templates
20,157SLO Classifications
18,440 Students
857 Teachers
117 Schools
SLO Classifications from other District SLO Templates

[1] We made these classifications based on content groupings of the SLO templates found on the district's SLO website: https://departments.dpsk12.org/are/slowiki/wiki/Home.aspx

In Figure 1 we make a distinction between students and SLO classifications. Note that in Subpopulation 1, there are almost twice as many SLO classifications (103,777) as there are students (55,479). This is because many teachers use the same set of students to enact the SLO process. For example, one class of students might have separate periods for math and English Language Arts (ELA) instruction, with each teacher establishing distinct math and ELA SLOs for the same students. Once we condition on SLOs that are based on ELA and Math Templates provided by the district, there is a much closer one-to-one relationship between students and SLO classifications, with the remaining difference attributable to students who appear to have been rated on the same SLO by multiple teachers.

SLOs based on English Language Arts (ELA) and Math district templates were clearly quite popular with teachers. Notice that among the 2,257 (1,400 + 857) teachers who chose to use district SLO templates for ELA or Math, SLOs for ELA (62%) were written considerably more frequently than SLOs for Math (38%).

DESCRIPTIVE STATISTICS

At the student level, there are three SLO-related variables of interest: 1) preparedness levels, 2) end-of-course command levels, and 3) growth points earned. Teachers assigned students to preparedness levels in the fall. They could choose from one of five preparedness levels: significantly underprepared, underprepared, somewhat prepared, prepared, or ahead. At the end of a course, teachers rated students' command (i.e., mastery) of the SLO. They could choose from one of five command levels: below limited command, limited command, moderate command, strong command, or distinguished command. Student growth scores were computed according to the relationship between a student's preparedness level and command level. With a few exceptions, these growth scores were computed automatically by the SLO application using a series of decision rules that were adopted by the district as part of the LEAP contract negotiations. The version of the decision rules shared with our team in September 2015 is presented in Figure 2. The grey boxes in Figure 2 reflect the decision points where manual entries are made by the evaluator to finalize a teacher rating.

FIGURE 2. DPS "SLO SCORING MATRIX"

	Below Limited Command	Limited Command	Moderate Command	Strong Command	Distinguished Command
Significantly Underprepared	Teacher & Evaluator Decision: 0, 1, or 2	3	3	3	3
Underprepared	Teacher & Evaluator Decision: 0 or 1	2	3	3	3
Somewhat Prepared	0	1	2	3	3
Prepared	NA*	0	1	2	3
Ahead	NA*	0	0	1	Teacher & Evaluator Decision: 2 or 3

Teacher & Evaluator Decision Cells:

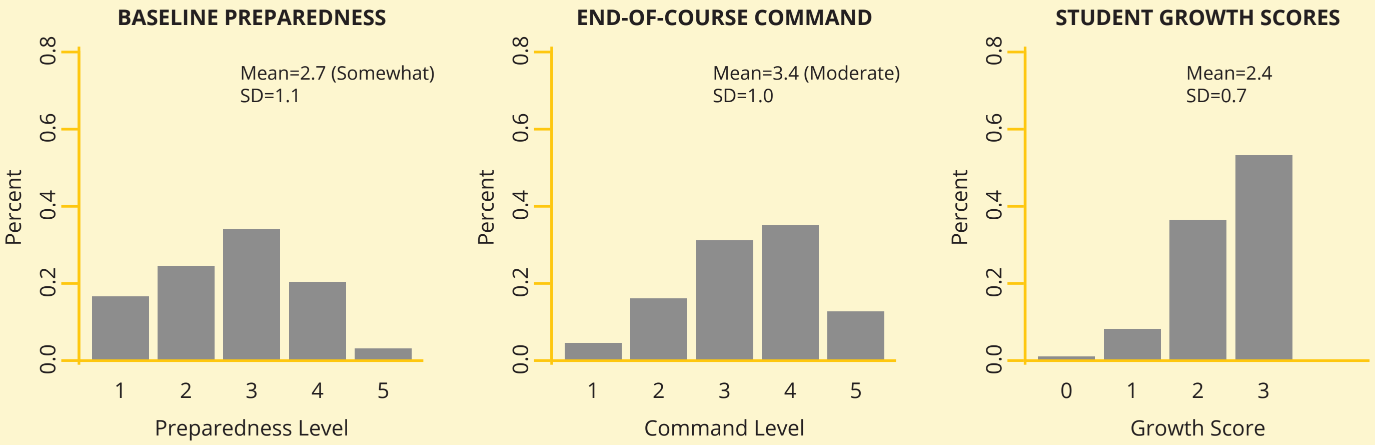
- *Growth can look different for individual students falling in these cells. For example, a Significantly Underprepared student can demonstrate substantial growth, but still not meet the criteria for Limited Command of the current year standards.*
- *In these cells, teachers and evaluators determine the student's growth based on the individual student's body of evidence.*

Possible Points = # students X 3

* For students starting a course Prepared or Ahead, Below Limited Command represents less mastery than they begin the course with. This is based on Limited Command representing a level of mastery expected around the beginning of a course. Thus, Below Limited Command is not an option for these students. Limited Command is the lowest level of mastery they can attain.

Figure 3 presents frequency distributions of these SLO variables from Subpopulation 1 using distinct SLO classifications as the unit of analysis. These results indicate that the average student was classified as "somewhat prepared" for their SLO at the start of the school year and as having "moderate command" of the SLO by the end of the instructional period. Most students tracked on SLOs received 2 or 3 growth points out of a maximum of 3 points. This suggests that, according to their teachers, the vast majority of students showed evidence of moderate to high growth.

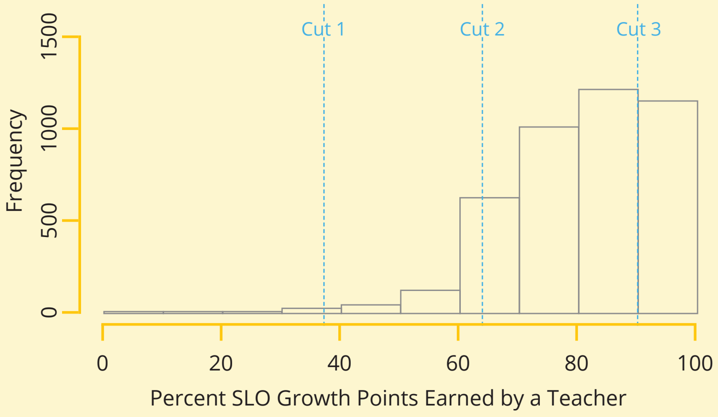
FIGURE 3. DISTRIBUTIONS OF SLO OUTCOMES FOR STUDENTS FROM TRADITIONAL SCHOOLS (N = 103,777).



The SLO variable of interest at the teacher level is the percent of SLO points earned by each teacher. We calculated the percent of SLO points earned by each teacher using the student growth points provided by the district. We followed the procedure listed below given by the LEAP SLO growth rules:

1. Sum the student growth points associated with each unique teacher ID.
2. Calculate the number of students a teacher rated on SLOs.
3. Calculate the maximum possible growth points a teacher could earn on an SLO by multiplying the number of students a teacher rated on SLOs by 3.
4. Calculate the percent of SLO points earned by a teacher by dividing #1 by #3.

FIGURE 4. DISTRIBUTION OF PERCENT OF SLO GROWTH POINTS EARNED BY A TEACHER.



We computed the percent of SLO points earned by a teacher for all teachers in the traditional schools subpopulation. On the basis of their percent of SLO points earned, teachers were assigned a growth classification based on cut scores set by the district. Out of 4,187 teachers in the subpopulation of traditional schools, just .2% were rated "ineffective," 7.1% of teachers were rated "approaching," 65.3% were rated "effective," and 27.4% were rated "distinguished." Figure 4 displays the distribution of this variable with the LEAP cutpoints superimposed.

The distribution of the percent of SLO growth points variable has a strong skew toward high values. To put this in perspective, one third of all teachers (1,376 out of 4,187) earned 90% or more of possible SLO growth points; one tenth of all teachers earned 100% of possible SLO growth points. This is indicative of a ceiling effect on teacher growth scores, and may be the result of SLOs that were not sufficiently ambitious (i.e., it was very easy for students to demonstrate strong or distinguished command) or of student growth scores that were inflated because students' levels of preparedness were set too low. We explore this second hypothesis in the next section.

VALIDITY OF PREPAREDNESS CLASSIFICATIONS

To examine the validity of teachers' SLO preparedness classifications, we restrict our attention to SLO variables from subpopulations 2 and 3 (ELA and Math SLOs). We used students' ELA and Math PARCC scores from the previous school year (spring of 2014-15) as a criterion to characterize the consistency of teachers' decisions. An important limitation of this analysis is that it could only be performed for students with ELA and/or Math SLO classifications who have PARCC scores from the preceding year.

We distinguish between SLO preparedness classifications that were consistent and those that were inconsistent by comparing students' ELA and Math SLO preparedness levels to their prior grade PARCC performance levels in the same test subject. Table 1 shows the cross-tabulation of these two student classifications by test subject.

TABLE 1. CROSSTAB OF SLO PREPAREDNESS LEVELS WITH PARCC PERFORMANCE LEVELS

PARCC Performance Level	SLO Preparedness Level					
	Sig. Up. [1]	Up. [2]	S. Prep. [3]	Prepared [4]	Ahead [5]	Totals
ENGLISH LANGUAGE ARTS						
Did not yet meet expectations [1]	1,344	780	402	42	2	2,570
Partially Met Expectations [2]	549	916	854	163	3	2,485
Approached Expectations [3]	247	657	1,088	561	43	2,596
Met Expectations [4]	86	363	1,065	1,064	211	2,789
Exceeded Expectations [5]	8	36	180	278	102	604
Total	2,234	2,752	3,589	2,108	361	11,044
MATHEMATICS						
Did not yet meet expectations [1]	743	655	380	64	2	1,844
Partially Met Expectations [2]	460	666	634	158	5	1,923
Approached Expectations [3]	267	654	850	404	30	2,205
Met Expectations [4]	104	425	1,001	911	184	2,625
Exceeded Expectations [5]	3	46	213	281	137	680
Total	1,577	2,446	3,078	1,818	358	9,277

Notes: Green cells are considered "consistent SLO classifications" and those that are shaded in blue are considered to be "inconsistent."

There are a number of different ways that one could characterize preparedness classifications that are consistent with PARCC test score performance. The strictest rule would be a one to one agreement between PARCC performance levels and SLO preparedness level such that only cells along the main diagonal of each cross-tabulation are considered to be consistent. Using this criterion, we find that 41% of ELA and 35% of Math SLO preparedness classifications were consistent.

However, there is no particular reason to require a one to one relationship between PARCC performance levels and SLO preparedness levels. An alternative approach would be to regard the PARCC performance levels of 3 (“approached expectations) and 4 (“met expectations) as a key dividing line for preparedness classifications. That is, if a student fell in a PARCC performance level of 4 or 5, we consider a preparedness level of 4 or 5 to be a consistent classification for that student (cells shaded green), and a preparedness level of 1, 2 and 3 to be an inconsistent classification (cells shaded blue). For a student in PARCC performance level of 3 or lower, so long as the student’s preparedness levels is also 3 or lower, it is considered to be a consistent classification (cells shaded green), while a preparedness level of 4 or 5 is considered an inconsistent classification. Using this criterion, we find that 77% of ELA and 74% of Math SLO preparedness classifications were consistent[2].

Table 1 indicates that at least some teachers may have had a tendency to underestimate the preparedness levels of their students. This is seen most clearly by looking at the 3,483 and 3,305 students who either met or exceeded expectations on the PARCC tests for ELA and Mathematics respectively (performance level 4 or 5). Approximately, 51% and 54% of these students were classified inconsistently in a downward direction (significantly underprepared, underprepared or only somewhat prepared). In contrast, consider the 7,651 and 5,972 students who were classified as approaching expectations or lower on the PARCC tests for ELA and Mathematics (performance level 1, 2 or 3). Only about 11% of students in these groups were classified inconsistently in an upward direction as prepared or ahead for their SLO. This demonstrates an asymmetry in inconsistent preparedness classifications—teachers were more likely to underestimate a student’s preparedness relative to PARCC performance than they were to overestimate it.

It is important to note that an inconsistent classification, based on the definitions illustrated in Table 1, are not necessarily inaccurate classifications. It is entirely possible that a student who performed at a high level on the prior grade PARCC test might not be viewed by a teacher as prepared for an SLO at the beginning of the year. Explanations for inconsistent classifications could be plausibly rooted in differences between the content of the prior grade PARCC test and the current year SLO, or might even be attributable to summer learning loss.

Next, we examine whether certain kinds of students were more likely than others to have their preparedness potentially underestimated. We do this by examining only those students[3] who scored in performance levels of 4 or 5 on PARCC math and ELA tests. Table 2 summarizes the characteristics of students in these two groups.

TABLE 2. DESCRIPTIVE STATISTICS FOR STUDENTS BY PARCC PERFORMANCE LEVEL CLASSIFICATIONS

	STUDENTS IN PARCC PERFORMANCE LEVELS 4 OR 5	
	MATH	ELA
% Female	56.0	55.2
% Nonwhite	50.2	51.4
% FRL	37.0	41.0
% ELL	23.9	26.5
% IEP	1.5	1.3
% GT	48.8	45.8
N	3300	3390

[2] As part of a meeting with ARE staff in May 2016, another permutation for consistent vs inconsistent classification was proposed in which students who had partially met or where approaching expectations on PARCC tests would be considered inconsistently classified if given a preparedness level of significantly underprepared or underprepared. Using this criterion, we would find that 64% of ELA and 59% of Math SLO preparedness classifications were consistent.

[3] The unit of analysis in these logistic regressions is actually a student preparedness classification.

We specify a unique logistic regression for ELA and Math SLO preparedness classifications, respectively. A value of 1 for the outcome variable indicates that a student was inconsistently classified in a manner that may have underestimated the student's level of preparedness, and a value of 0 indicates that the student was consistently classified into a preparedness level of either 4 or 5. The predictor variables in the logistic regression are *Female, Nonwhite, FRL, ELL, IEP,* and *GT*. Each represents an student characteristic variable that takes on a value of 1 if a student is female, nonwhite, eligible for free or reduced price lunches (FRL), an English Language Learner (ELL), receiving special education services through an individualized education plan (IEP), or part of the gifted and talented (GT) program, and a value of 0 otherwise. In Table 3 below, we provide a summary of the key results (more detailed results are provided in the appendix).

TABLE 3. RESULTS FROM LOGISTIC REGRESSIONS USING STUDENTS WITH 2014-15 PARCC PROFICIENCY LEVELS OF 4 OR 5

	MATH SLO	ELA SLO
Baseline (Unconditional) Probability of Inconsistent Classification [Preparedness Potentially Underestimated]	.542	.512
Change in Probability of Inconsistent Classification if Student is		
Female	.046*	-.052**
Nonwhite	.091***	.028
FRL Eligible	.124***	.025
ELL	.050	.040
IEP	.044	.134
Gifted & Talented	-.118***	-.101***

*Note: * p < .05, ** p < .01, *** p < .001*

For Math SLOs, female students, nonwhite students and those eligible for FRL services are significantly more likely to have their preparedness underestimated relative to their male, white and non-FRL eligible peers. The values in the table indicate the marginal amount that a student's probability of being inconsistently classified would increase (positive value) or decrease (negative value) as a function of their personal characteristics. For example, The values of .046 and -.052 for the *Female* predictor variable indicate that for math SLOs, females were 4.6% more likely to have their preparedness inconsistently classified relative to males, while for ELA SLOs, females were 5.2% less likely to be inconsistently classified.

To better appreciate the implications of these results, consider two DPS students with Math SLOs that are both native English speakers and have not been provided with either special education or gifted and talented designations. The first student is male, white and non-FRL eligible; the second student is female, nonwhite and FRL eligible. The first student would have about a 48% chance of having his SLO preparedness level underestimated, but for the second student, the probability of underestimation jumps by 24% up to a 72% chance. Fortunately, differences in preparedness classifications of this magnitude were only evident for math SLOs, not for ELA SLOs.

Note that for both Math and ELA SLOs, students flagged for gifted and talented status have about an 11% and 10% *lower* chance of having their preparedness classified inconsistently. This comes as little surprise since teachers are likely to be aware of a student's gifted and talented status when they are making preparedness classifications.

In summary, these results suggest that teachers may be systematically underestimating students' preparedness levels, and with respect to preparedness for Math SLOs, they appear to be underestimating the preparedness of certain kinds of students more than others. The general finding that many students had preparedness classifications that were lower than what was suggested by their PARCC performance has a number of possible explanations. One explanation, consistent with the qualitative findings described in Brief 2, is that many teachers had the (mistaken) impression that SLO preparedness and command exist on the same dimension. A teacher might look past the labels used for the preparedness and command categories and instead focus just on the numbers. From this perspective, it might be hard to imagine that a student at the beginning of the year could be at level 4 on a scale that runs from 1 to 5. Another explanation would be that teachers, aware that student growth would count toward their LEAP evaluation ratings, responded to this incentive by placing students into lower preparedness to artificially inflate their growth ratings. Of course, these two explanations are not mutually exclusive.

VALIDITY OF TEACHER SLO GROWTH POINTS

In this section, we shift from examining the validity of student-level preparedness classifications to the validity of using SLO scores as a measure of teacher effectiveness. We do so by adopting a framework that has been used to evaluate the properties of value-added models in the context of teachers who teach students for whom state-administered standardized test scores are available across adjacent years. When averaged over students and attached to teachers, an SLO measure can be cast as a crude version of a value-added model in that it attempts to distinguish teachers on the basis of differences in mastery conditional on their incoming levels of preparedness. In a typical value-added model for teachers in tested subjects, preparedness and mastery would be established objectively on the basis of prior and current grade standardized tests; on SLOs, both of these scores are determined somewhat subjectively by teachers. Because value-added models take pre-existing differences in student achievement (and often other variables) into account, they provide a fairer basis for comparing teachers relative to comparisons based solely on students' end of year achievement. With this in mind, we can examine to what extent a teacher-level SLO measure levels the playing field in a manner analogous to a value-added model.

We examine ELA and Math teachers from subpopulations 2 and 3 in this analysis. In what follows, we contrast two different teacher-level SLO outcome variables. The first, "AvgEOY" is the average of student's end-of-year SLO mastery classification (on a scale from 1 to 5). The second, "PctSLOPts" is the percent of possible growth points earned by a teacher (on a scale from 0% to 100%). Table 4 shows correlations between these two variables as well as between other available and relevant teacher-level variables. Notably, this includes the professional practice scores of teachers that derive from DPS's LEAP professional practice rating ("1415LEAP.PPPts"). A teacher's overall professional practice rating is a combination of classroom observation ratings, a professionalism rating, and student perception survey ratings. Teachers' professional practice scores are normally distributed with a mean of 37.1 and standard deviation of 3.4.

TABLE 4. RELATIONSHIPS AMONG TEACHER-LEVEL VARIABLES

	AVG EOY	PCT SLOPTS	1415 LEAP PPPts	% GT	% IEP	% FRL	% ELL	PARCC. ELA	PARCC. Math
AvgEOY	(0.67)								
PctSLOPts	0.48	(13.6)							
1415LEAP.PPPts	0.13	0.17	(3.4)						
%GT	0.20	-0.05	0.03	(19)					
%IEP	-0.38	0.02	-0.02	-0.23	(27)				
%FRL	-0.38	-0.03	-0.19	-0.42	0.18	(31)			
%ELL	-0.34	-0.13	-0.15	-0.11	0.03	0.58	(32)		
PARCC.ELA	0.59	0.03	0.10	0.73	-0.58	-0.70	-0.44	(28.6)	
PARCC.Math	0.65	0.06	0.12	0.71	-0.50	-0.73	-0.46	0.89	(24.8)

Note: Correlations were computed using pairwise complete observations. The values in parentheses along the main diagonal represent the standard deviation of each variable.

Notice in Table 4 that AvgEOY tends to be more strongly correlated with aggregated demographic and achievement variables than PctSLOPts. In particular, while average PARCC scale scores in ELA and Math have a correlation of 0.59 and 0.65 with AvgEOY, the two variables are essentially uncorrelated with PctSLOPts. Interestingly, the only teacher-level variable for which the correlation does not drop when going from AvgEOY to PctSLOPts is the professional practice measure 1415LEAP.PPPts.

We explore these findings more formally by regressing the two teacher SLO outcome variables on a set of variables that capture aggregate characteristics of the students in each teacher's class, the grade level being taught, and the ratings of professional practice.

$$\begin{aligned} \mathit{Tch.SLO.Outcome} & \\ &= \beta_0 + \beta_1\,\mathit{Female} + \beta_2\,\mathit{Nonwhite} + \beta_3\,\mathit{Elementary} + \beta_4\,\mathit{Middle} + \beta_5\,\mathit{High} \\ &+ \beta_6\%\mathit{GT} + \beta_7\%\mathit{IEP} + \beta_8\%\mathit{FRL} + \beta_9\%\mathit{ELL} + \beta_{10}\ \ \mathit{1415\_PPPts} \\ &+ \beta_{11}\,\mathit{PARCC.SS.ELA} + \beta_{12}\,\mathit{PARCC.SS.Math} + \varepsilon \end{aligned}$$

Tch.SLO.Outcome is a teacher's average end-of-year SLO classification or a teacher's percent of SLO points earned. *Female* and *Nonwhite* are dummy variables that indicate whether a teacher is female or nonwhite. *Elementary, Middle*, and *High* are dummy variables that indicate the grade level of a teacher. *%GT, %IEP, %FRL*, and *%ELL* are all teacher-level variables that were created by aggregating student demographic information for a particular teacher using just the students a teacher tracked on SLOs. 1415*_PPPts* are the amount of professional practice points a teacher earned in 2014-15 under the LEAP teacher evaluation system. *PARCC.SS.ELA* and *PARCC.SS.Math* are aggregated student PARCC scaled scores from spring 2014-15 using just the students a teacher tracked on SLOs. This variable was standardized within subject. Because very few teachers taught both ELA and Math and submitted SLOs for both of these subjects, we ran two separate regressions, one for each subject. The estimated coefficients are presented in Table 5.

TABLE 5. REGRESSION MODELS FOR TEACHERS' SLO OUTCOMES.

	Average End-of-Year SLO Classifications		Percent of SLO Growth Points Earned	
	ELA	Math	ELA	Math
Intercept	2.67***	2.30***	51.60***	56.05***
Female	0.06	0.13*	-0.73	1.98
NonWhite	-0.12	0.12	-3.33	0.33
Elementary	0.11	0.09	2.56	-1.33
Middle	0.01	-0.09	1.37	-0.34
High	-0.04	0.01	2.49	0.25
% GT	0.23	0.0018	-0.0004	0.0003
% IEP	-0.0069***	-0.0049***	0.0001	0.0000
% FRL	-0.0010	-0.0051*	0.0000	-0.0003
% ELL	-0.0040**	0.0022	-0.0001	0.0005
1415_PPPts	0.01	0.01	0.71**	0.57*
PARCC.SS.ELA	0.16*		1.05	
PARCC.SS.Math		0.27***		-0.29
Deg. of Freedom	327	248	327	248
Adj. R-Squared	0.4353	0.5492	0.04588	0.04856

***** *$p<0.001$, ** $p<0.01$, * $p<0.05$*

Teachers' SLO outcomes related to status (average end-of-year SLO classifications) are associated with aggregate student characteristics such as %IEP, %FRL, and %ELL, though this varies somewhat by SLO subject. For both ELA and Math, the more IEP students taught by a teacher, the lower that teacher's average end-of-year classifications. For ELA teachers, larger proportions of ELL students taught by a teacher are also associated with lower average end-of-year classifications. For math teachers, larger proportions of FRL students taught by a teacher are also associated with lower average end-of-year classifications. Also, female math teachers are associated with a small average increase in average end-of-year classifications compared to non-female teachers. These results suggest that the average end-of-year SLO mastery classifications are associated with the demographic of a teacher and their students.

The story is quite different for the SLO growth measure. The only statistically significant covariate when the outcome shifts to percent of SLO growth points is teachers' LEAP professional practice measure. For ELA and math, a one point increase in a teacher's professional practice score is on average associated with a 0.71 and 0.57 increase in SLO growth points, respectively. Expressed in effect size units (standard deviations of the outcome), a 1 standard deviation increase in professional practice points is associated with a 0.17 standard deviation increase in percent of SLO growth points earned.

This association between SLO growth and teachers' professional practice is both statistically and practically significant. Teachers' percent of SLO growth points earned appears to be capturing information about teachers that has some association with the professional practice measure from LEAP. Just as importantly, we see that teachers' growth points are not associated with measurable classroom contexts. This suggests that the aggregation of students' growth into a teacher-level measure is more indicative of salient differences in a teacher's effectiveness than to student characteristics like GT, IEP, FRL, or ELL status that are outside a teacher's control.

These findings put the use of SLOs as a teacher-level growth measure in a surprisingly positive light. Even though there is evidence that the SLO growth scores of students were biased upwards by the way that teachers placed them into preparedness categories, we see no evidence that this bias has a differential impact on teacher-level measures. However, we must emphasize that these findings should be regarded as tentative and require further investigation. In particular, we note that our available covariates are only able to explain about 5% of the variance in the percent of SLO growth points earned outcome. It is also important to note that our findings here are limited to teachers with students for whom PARCC test scores were available, so they may not generalize to the full population of DPS teachers.

The release of PARCC results for 2015-16 should open the door to a number of additional analyses. An important comparison of interest will be between teacher-level SLO measures and mean student growth percentiles.

RECOMMENDATIONS

- Given the popularity of SLO templates for math and ELA, the district should continue to improve the quality of these templates and accompanying resources.
- The district should take steps to reduce the likelihood of student growth scores on SLOs that are inflated because teachers have set their level of preparedness too low.
- The pre-existing training and guidance regarding the process of classifying students into preparedness levels may need to be improved.
- Principals and teacher leaders should work with teachers who choose to write Math SLOs to reduce the likelihood that they underestimate the preparedness of female, nonwhite and FRL-eligible students.
- Additional analyses should be conducted to compare the properties of teacher-level SLO measures to teacher-level student growth percentile aggregates, and to examine the year to year stability of SLO measures.

APPENDIX

TABLE A

Results from English Language Arts logistic regression using students with 2014-15 PARCC proficiency levels of 4 or 5

	Log Odds	Std. Error	z-value	Pr(>| z|)	Odds Ratio
Intercept	-0.20**	0.07	-2.79	0.005	0.82
Female	0.21**	0.07	2.96	0.003	1.23
Non-White	-0.11	0.09	-1.26	0.210	0.89
FRL	-0.10	0.09	-1.10	0.273	0.90
ELL	-0.16	0.10	-1.68	0.093	0.85
IEP	-0.57	0.32	-1.81	0.071	0.56
GT	0.41***	0.07	5.79	7.050 E-09	1.50

**** p<0.001, ** p<0.01, * p<0.05*

TABLE B

Results from math logistic regression using students with 2014-15 PARCC proficiency levels of 4 or 5

	Log Odds	Std. Error	z-value	Pr(>| z|)	Odds Ratio
Intercept	0.10	0.07	1.35	0.18	1.11
Female	-0.18*	0.07	-2.51	0.01	0.83
Non-White	-0.37***	0.09	-4.10	4.22 E-05	0.69
FRL	-0.50***	0.10	-5.19	2.11 E-07	0.61
ELL	-0.20	0.11	-1.90	0.06	0.82
IEP	-0.18	0.30	-0.58	0.56	0.84
GT	0.49***	0.07	6.74	1.54 E-11	1.63

**** p<0.001, ** p<0.01, * p<0.05*

TABLE C

Comparison of deviance statistics from ELA and Math logistic regressions

	Null deviance (degrees of freedom)	Residual deviance (degrees of freedom)	AIC
ELA	4,697.5 (3,389)	4,630.9 (3,383)	4,644.9
Math	4,551.0 (3,299)	4,341.6 (3,293)	4,355.6