Examining the Dual Purpose Use of Student Learning Objectives for Classroom Assessment and Teacher Evaluation

Derek C. Briggs

Rajendra Chattergoon

Amy Burkhardt

University of Colorado Boulder

June 15, 2018

Examining the Dual Purpose Use of Student Learning Objectives for Classroom Assessment and Teacher Evaluation

Abstract

The process of setting and evaluating *Student Learning Objectives* (SLOs) has become increasingly popular as an example where classroom assessment is intended to fulfill the dual purpose use of informing instruction and holding teachers accountable. A concern is that the high stakes purpose may lead to distortions in the inferences about students and teachers that SLOs can support. This concern is explored in the present study by contrasting student SLO scores in a large urban school district to performance on a common objective external criterion. This external criterion is used to evaluate the extent to which student growth scores appear to be inflated. Using two years of data, growth comparisons are also made at the teacher-level for teachers who submit SLOs and have students that take the state-administered large-scale assessment. Although they do show similar relationships with demographic covariates and have the same degree of stability across years, the two different measures of growth are weakly correlated.

Introduction

Classroom assessment and large-scale assessment are generally understood to serve different purposes. To the extent that there is a conventional distinction to be drawn between the two, it is that classroom assessments are optimal as a means of conveying feedback about student learning (i.e., assessment *for* learning), while large-scale assessments are optimal as a means of conveying information that can be used to monitor and evaluate student learning (i.e., assessment *of* learning). This distinction seems sensible since classroom assessments tend to be locally developed, their administration is in the control of teachers, and they are generally proximal to the teacher and students' enacted curriculum (c.f., Ruiz-Primo, Shavelson, Hamilton & Klein, 2002). Furthermore, the feedback students and teachers receive from these assessments is often immediate or follows from a short time lag. In contrast, large-scale assessments are externally developed, their administration is outside the control of teachers, and the results have—historically at least—not been available for many months. However, because of their standardization, they provide for an efficient and reliable way to compare students across classrooms, schools and districts.

While these conventional distinctions in the purposes served by classroom and large-scale assessments might appear self-evident, an increase in educational accountability policies that emphasize teacher evaluation has led to some blurring of boundaries. In many instances, large-scale assessments have been increasingly expected to not only provide information relevant to systems of educational accountability, but to also provide information that is formatively useful to teachers, students and parents. One reason for this is the potential for computerized adaptive assessments to provide teachers with much more timely (even immediate) feedback on student performance relative to traditional paper and pencil exams[1]. This boundary blurring between assessments used for both high and low-stakes purposes is also becoming more prevalent coming from the opposite direction. That is, in some states and school districts, locally developed assessments, previously used solely to inform instruction or support student grading within classrooms, are being used for comparative purposes *across* classrooms. One of the highest profile examples of this can be found in the process commonly used to set and evaluate *Student Learning Objectives* (SLOs).

Our focus in the present study is on the use of SLOs in support of teacher evaluation *and* formative assessment practices in Denver Public Schools (DPS). On the one hand, there is a clear desire among DPS assessment and evaluation staff for SLOs to be used formatively to inform instructional practice. On the other hand, DPS teachers are well aware that SLOs also provide the district with aggregate measures that can count both toward a teacher's monetary compensation and their annual evaluation. The motivating question we explore in this study is whether there is evidence that the use of classroom assessment for high-stakes purposes vis-à-vis their role in SLOs is associated with distortions in the information the assessments convey about student growth.

[1] The popularity of computer-based "interim" assessments in many school districts across the country suggests that many educators believe that these products are formatively useful, even though the assessments are standardized and externally developed. Yet because these sorts of assessments have generally not been used as inputs for high-stakes accountability decisions, their "dual use" potential is an open question.

Student Learning Objectives

Over the past decade, a growing number of states and school districts in the United States have sought to incorporate evidence of student growth into formal evaluations of teachers and schools under the auspices of educational accountability (c.f., Dougherty & Jacobs, 2013). Although a great deal of research and debate has surrounded the use of statistical models for this purpose, only about one third of classroom teachers teach students for whom state-administered standardized tests are available as inputs (Hall, Gagnon, Schneider, Marion, Thompson, 2014). Hence, for a majority of teachers, other evidence is needed to support inferences about student growth. In more than 30 states, this evidence has been, and/or continues to be, gathered through the development and evaluation of student growth using SLOs (Lacireno-Paquet, Morgan & Mello, 2014). Indeed, in many places SLOs are now being developed and submitted by teachers in both state-tested- and non-state-tested subjects and grades (Hall, et al., 2014). Common elements of the SLO process typically include the following:

- specification of a goal as part of an "objective statement" that defines the content and skills students are expected to learn;
- specification of the interval of instruction over which the learning is expected to occur (e.g., semester or year-long).
- identification of assessments to be given to students at the beginning and end of the instructional interval.
- specification of growth targets set for individual students based on performance achieved on identified assessments.

- designation of a teacher rating based on growth achieved by students.

The implementation of these common SLO elements varies across states and districts depending upon the degree of centralization and comparability desired in the set of learning goals used across all teachers, the set of data sources used to support the process, and the methodology used to compute teacher ratings (Lachlan-Hache, Bivona, Reese, Cushing, & Mean, 2013).

In theory, SLOs can provide a means for teachers to establish learning goals, monitor students' progress toward these goals, and then evaluate the degree to which students achieve these goals using relevant measures. In this sense, SLOs merely represent the formalization of a process that should be part of good teaching. What makes them somewhat unique is that beyond their instructional use, SLOs are intended to provide a quantitative source of evidence that can be used to differentiate effective and ineffective teachers. As a quantitative indicator, this brings SLOs squarely into the realm of what has been referred to as Campbell's Law (Campbell, 1976; Briggs, 2016).

> The more any quantitative social indicator (or even some qualitative indicator) is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell, 1976, p. 49).

SLOs are especially susceptible to corruption because it will often be the case (as it is in the data context in this study) that many, or even all, of the elements that figure into the quantification and aggregation of student growth—the choice of learning objective, the choice of assessments to administer, the scoring of these assessments, and the classification of students into mastery levels—are directly influenced by teachers' judgments.

Student Learning Objectives in Denver Public Schools

Denver Public Schools was one of the first school districts in the country to establish a formal process for using locally developed assessments to serve both instructional and evaluative purposes (Gonring, Teske, & Jupp, 2007; Hershberg & Robertson-Kraft, 2009). A pilot version of SLOs (called "Student Growth Objectives") was being used by subsets of DPS schools as far back as 2001, and by 2006, all DPS teachers received small salary bonuses or salary increases of up to $376 if some designated percentage of students could be shown to have met one or two of their growth objectives. More recently, as of the 2015-16 school year, results from SLOs have been factored into teacher evaluation. It can be argued that the dual use concept for SLOs was pioneered by DPS, and that the SLO process taken in many other districts and states across the country represents an emulation of the DPS approach.

The Role SLOs Play in the DPS Accountability Context

By state law[2], public school teachers in Colorado must be evaluated annually, and 50% of this evaluation must be based on evidence of student growth in academic achievement. In DPS, 50% of a teacher's overall Leading Effective Academic Practice

[2] https://www.cde.state.co.us/educatoreffectiveness/senatebill10191rulesdocument

(LEAP) [3] evaluation score comes from three sources: a collective measure of school growth, student growth on state-administered achievement tests, and student growth as indicated by SLO performance. Among these three sources, a measure of school-level growth contributes 10%, and another 10% comes from the mean of student growth percentiles (Betebenner, 2009) computed for students who have taken the state-administered tests (i.e., the Partnership for Assessment of Readiness in College and Career Consortium tests [PARCC]) in English Language Arts (ELA) and/or Math two years in a row. For these same teachers, SLOs contribute the remaining 30% of growth evidence toward a LEAP rating. For teachers that do not have students who take state-administered tests, SLOs contribute 40% to their overall LEAP rating. Hence, for all teachers, whether their students take state-administered large-scale assessments or not, student performance on SLOs is by far the predominant source of information used to satisfy the growth component of the evaluation system under Colorado state law.

Components of SLOs at Denver Public Schools

In this study we focus on DPS SLO data from the 2015-16 and 2016-17 school years. DPS's Department of Accountability, Research and Evaluation offered the following description of SLOs in the 2016 edition of its "SLO Teacher Handbook" under the heading, "What are Student Learning Objectives?":

[3] http://careers.dpsk12.org/wp-content/uploads/2016/09/2017-LEAP-Teacher-Handbook-lo-res.pdf

> Effective teachers have learning goals for their students and use assessments to measure progress toward these goals. They have a deep understanding of where students are at the beginning of a course, and what they can achieve by the end. Effective teachers analyze standards, select and administer rigorous assessments aligned to those standards, and measure how their students grow during the school year. They use this data to drive their instruction and are constantly reflecting on and refining their craft. Student Learning Objectives (SLOs) embody these effective pedagogical practices by helping DPS educators focus on high impact standards, set ambitious learning goals, and measure students' progress toward attaining them. This process will yield greater student growth on critical learning outcomes by allowing teachers to plan backward from an end vision of student success, ensuring that every minute of instruction is geared toward our district vision that *Every Child Succeeds*.

This description makes the intended formative use of SLOs fairly clear. Beyond the evaluative role they play as an input into a teacher's LEAP rating, SLOs are expected to provide a tool for teachers to use to facilitate student achievement and to improve upon their pedagogical practices.

Any given SLO in DPS contains three key components: an objective statement, performance criteria, and a "learning progression" rubric. Although teachers can create SLOs on their own, and many do, DPS staff provide two "template" SLOs in math and ELA for each grade level from Kindergarten through the 12th grade. To provide a concrete example, we can use the example of a district-developed grade 2 math SLO template for understanding the concept of place value. The objective statement of this SLO is

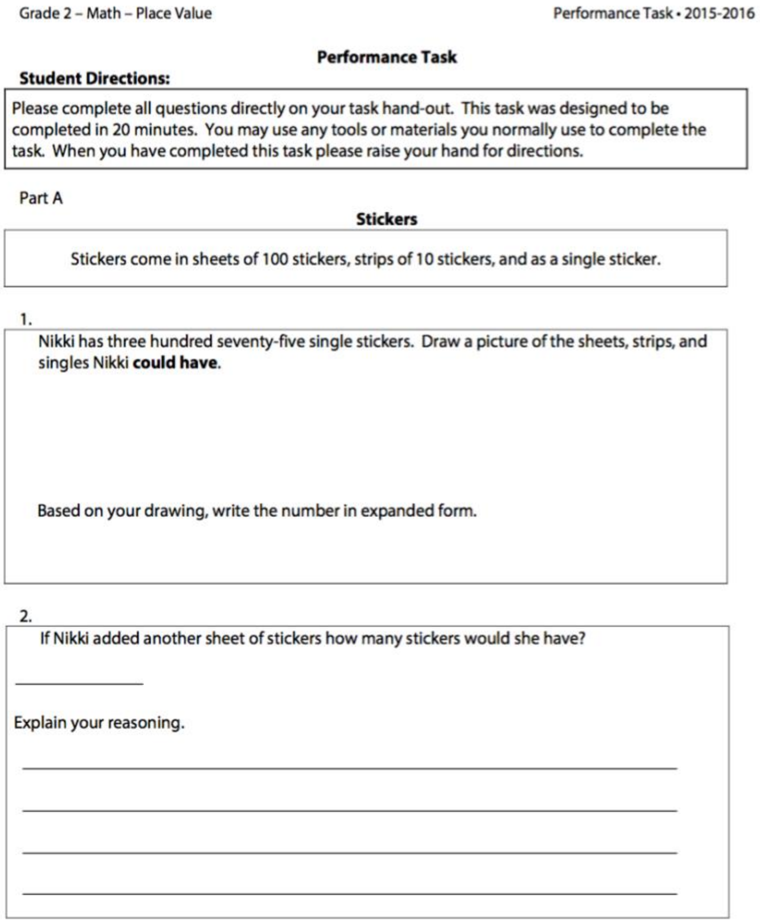
> All students will be able to model and explain the meaning of the digits in three-digit numbers as the amount of hundreds, tens, and ones both verbally and in writing. Students will apply this understanding to compare two three-digit numbers, use the symbols >, <, and = to record the comparison results, and justify the comparison both verbally and in writing.

The four performance criteria used to operationalize this objective are

1. Students demonstrate understanding of the three-digits of a three-digit number by independently modeling a three-digit number with a visual representation.
2. Students compare two three-digit numbers and record the results using symbols >, < and =.
3. Students independently express a three-digit number in expanded notation. Students represent the quantity in terms of hundreds, tens, and ones.
4. Students use grade-level academic and content language to explain their representations and justify comparisons orally and in writing based on the place value system.

Finally, a rubric is used to score/categorize students with respect to each of these performance criteria on a criterion-referenced scale from 1 to 4. Along with these three components, each DPS template comes with one or more performance-based assessments that teachers are encouraged to administer near the end of the SLO instructional period. Two of the six items for a grade 2 place value assessment are depicted in Figure 1. This assessment was designed under the leadership of content experts in math found in the district's curriculum and instruction division, and was written to align to its associated SLO objective statements and performance criteria, which are themselves aligned to the Common Core of State Standards.

Figure 1. District-developed Assessment Tasks for Place Value SLO



Grade 2 – Math – Place Value

Performance Task • 2015-2016

Performance Task

Student Directions:

Please complete all questions directly on your task hand-out. This task was designed to be completed in 20 minutes. You may use any tools or materials you normally use to complete the task. When you have completed this task please raise your hand for directions.

Part A

Stickers

Stickers come in sheets of 100 stickers, strips of 10 stickers, and as a single sticker.

1\.

Nikki has three hundred seventy-five single stickers. Draw a picture of the sheets, strips, and singles Nikki **could have**.

Based on your drawing, write the number in expanded form.

2\.

If Nikki added another sheet of stickers how many stickers would she have?

____

Explain your reasoning.

____

____

____

____

Monitoring and Scoring of Students on SLOs

At the student level, there are three important SLO-related variables: 1) preparedness levels, 2) end-of-year command levels, and 3) growth points earned. DPS teachers assign students to preparedness levels at the outset of the instructional period for an SLO (typically the beginning of the academic school year in September). They are able to choose from one of five preparedness levels:

Significantly Underprepared [1]: Students who enter the course/grade with particularly minimal mastery of the prerequisite knowledge and skills for the course/grade.

Underprepared [2]: Students who enter the course/grade with minimal mastery of the prerequisite knowledge and skills for the course/grade.

Somewhat Prepared [3]: Students who enter the course/grade has some, but not all, prerequisite knowledge and skills for the course/grade

Prepared [4]: Students who enter the course/grade with sufficient prerequisite knowledge and skills for the course/grade. Students are academically prepared to engage in the content area of the SLO.

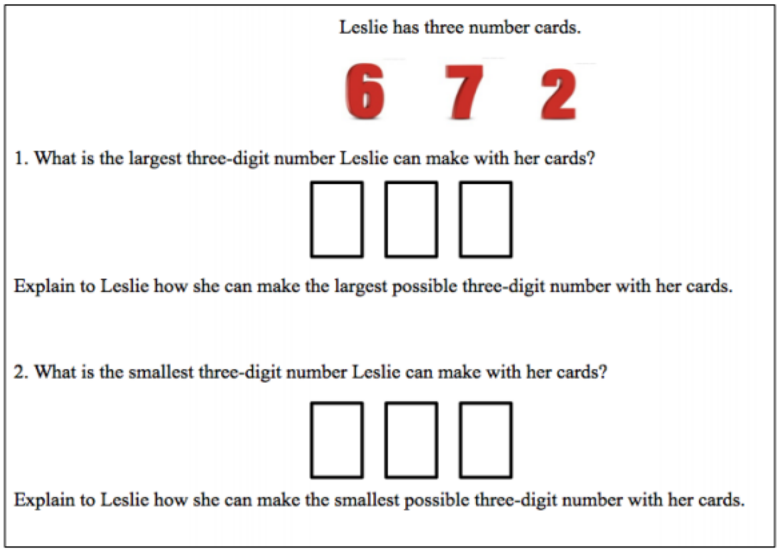
Ahead [5]: Students who enter the course/grade with a deep command of the prerequisite knowledge and skills for the course/grade. These students are able to apply previous learning to a variety of contexts.

At the end of a school year, teachers rate students' command of the SLO. They can choose from one of five command levels: below limited command, limited command, moderate command, strong command, or distinguished command[4]. In between the beginning and end of the school year, teachers are expected to monitor their students' progress, and to this end teachers can use one or more of the assessments provided by the district (if using a district template SLO, see example in Figure 1) and/or they can administer their own assessment tasks. Figure 2 provides an example of two assessment items for place value developed by a group of grade 2 teachers at a specific DPS elementary school. These particular teachers met regularly as part of a professional

[4] These were the descriptors used during the 2015-16 school year to align with the PARCC performance descriptors. For the 2016-17 school year these were changed to stay consistent with the change made to PARCC performance descriptors in the same year: (1) Did Not Yet Meet Expectations, (2) Partially Met Expectations, (3) Approached Expectations, (4) Met Expectations, (5) Exceeded Expectations.

learning community and were encouraged to share examples of their students' responses to progress monitoring tasks such as the ones illustrated in Figures 1 and 2. The intent was for discussions of these assessment results to prompt suggestions and strategies that could help teachers improve their instructional practices (for details, see Authors, 2014). However, these activities are not mandated or standardized, so the extent to which they occur will vary from school to school.

Figure 2. Example of a Locally Developed Assessment for Monitoring Understanding of Place Value



Leslie has three number cards.

6 7 2

1. What is the largest three-digit number Leslie can make with her cards?

□ □ □

Explain to Leslie how she can make the largest possible three-digit number with her cards.

2. What is the smallest three-digit number Leslie can make with her cards?

□ □ □

Explain to Leslie how she can make the smallest possible three-digit number with her cards.

In classifying their students both in terms of preparedness for the SLO at the start of the year, and level of command by the end of the year, teachers are instructed to use their professional judgment on the basis of the "body of evidence" collected for each student. Teachers are given considerable autonomy in classifying students into preparedness and command categories. For preparedness classifications, these measures are assembled primarily from student performance in prior year classes and assessments. For end-of-year command classifications, a body of evidence is defined formally by DPS as “data derived from a variety of assessment tools that measure the degree to which students are progressing toward each Performance Criterion, and more broadly the Objective Statement.” For students who use one of the district template SLOs, results from performance-based items (such as those illustrated in Figure 1) would be expected to play a prominent role, but these could also be supplemented with other teacher-developed assessments (such as the one illustrated in Figure 2). With regard to both classification decisions, teachers are expected to be able to provide a “strong, clear and thorough rationale” for the evidence used to support the classification, and each teacher’s SLO must be approved by their principal. However, beyond these guiding principles there is little standardization in the specific process that teachers use to make SLO classifications.

Student growth scores are computed according to the relationship between a student’s preparedness level and end-of-year command levels. With a few exceptions, these growth scores are computed automatically by an online SLO application using a series of decision rules. The version of the decision rules used during the 2015-16 school year is shown in Figure 3. (The grey boxes in Figure 3 reflect the decision points where manual entries are made by the evaluator to finalize a teacher rating.)

Figure 3. DPS "SLO Scoring Matrix"

	Below Limited Command	Limited Command	Moderate Command	Strong Command	Distinguished Command
Significantly Underprepared	0, 1, or 2	3	3	3	3
Underprepared	0 or 2	2	3	3	3
Somewhat Prepared	0	1	2	3	3
Prepared	NA*	0	1	2	3
Ahead	NA*	0	0	1	2 or 3

Note: Cells with gray shading reflect the decision points where manual entries are made by the evaluator to finalize a teacher rating.

* 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.

Finally, after SLO growth points have been computed for each student, a teacher-level SLO score is computed as the percent of maximum possible growth points earned by a teacher's students, where this maximum represents the number of students in the teacher's class multiplied by 3 (since each student can theoretically earn up to 3 points for their end of course level of command, regardless of their rated level of preparedness).

Some Threats to the Validity of SLOs as an Indicator of Growth

The SLO Scoring Matrix depicted in Figure 3 represents an example of a "value table" (Castellano & Ho, 2013) in which a student's end of course achievement status is differentially valued as a function of his or her incoming starting point or preparedness level. Because of this, a student that starts the year "significantly underprepared" or "underprepared" but ends at a "moderate" level of command for the SLO, can earn the same three points for growth as a student that starts the year "prepared" and ends with a "distinguished" level of command. The intent is to give both students and their teacher credit for the progress made during the school year, as opposed to potentially penalizing them for a lack of opportunity in the past.

In this study we examine a fundamental threat to the validity of SLOs as teacher-level measures of student growth: the potential for distortion caused by Campbell's Law. Unlike student growth percentiles (SGPs), which are based on objective and standardized measures of student achievement, SLOs are based upon teachers' holistic classifications of student achievement. Because these scores can represent 30 to 40% of a teacher's LEAP rating, this could create an incentive for teachers to inflate student growth either by classifying students as less "prepared" for an SLO than they actually are at the outset of an instructional period, or by classifying students as having a higher level of command than they have actually attained at the end of an instructional period. We look for correlational evidence that this kind of score distortion is occurring by using information about prior and current year student performance on Colorado's state-administered assessment (i.e., PARCC tests in ELA and Math at the time of this study) as a criterion against which the SLO preparedness of students can be compared. We examine whether students who scored in the "met expectations" or "exceeds expectations" performance levels on a PARCC test in ELA or Math 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 an SLO preparedness classification as evidence of a classification that appears to be inconsistent. We then more formally examine whether certain student characteristics are associated with the probability of a higher preparedness classification after controlling for prior year PARCC performance. We take a parallel approach with respect to end-of-year SLO command, but in this case we look for evidence that these scores have been inflated. Finally, we use a regression model to examine the extent to which a teacher-level SLO measure is associated with the aggregate characteristics of a teacher's students. We then contrast the use of a teacher-level measure of student growth based on SLOs to one based on to the use of SGPs (the mean SGP, or MGP).

Data and Descriptive Statistics

Our populations in 2015-16 and 2016-17 consist of all traditional district schools and teachers within those schools that implemented the SLO process. Alternative schools are excluded along with individual students who were not assigned SLO growth ratings due to factors such as attendance. One important distinction between the data collected in these two years is that 2015-16 represented the first year in which SLOs were fully implemented as a component of teachers' LEAP evaluation throughout all schools in DPS. However, in this first year of full implementation teachers were only required to submit one SLO. In the second year of full implementation, 2016-17, teachers were required to submit two SLOs. Another important distinction between the first and second year of this study is in the availability of prior grade PARCC scores to inform teacher preparedness ratings. Due to a delay in the release of these scores from the 2014-15 administration, prior grade scores were not available to DPS teachers in the fall of 2015 when they were making preparedness classifications for the 2015-16 school year. In contrast, prior grade PARCC scores were available for teachers to consult in the fall of 2016. This distinction likely explains one of the findings we present later, and we will return to it then.

As shown in Table 1, there were a total of 132 schools, 4,187 teachers and 55,479 unique students with SLO ratings in 2015-16. The following year, in 2016-17, there were a total of 140 schools, 4,383 teachers and 61,023 students. We further restrict the population in each year to only those teachers who chose SLOs that came from district-developed templates in the subjects of ELA and Math. We do this for two reasons. First, it allows for a contrast within the same subject domain relative to student performance on the PARCC assessment. Second, it reduces some of the variance in SLO outcomes from the inclusion of SLOs that may vary tremendously in quality. Filtering on the use of district-developed SLOs throughout ensures at least some minimum level of comparability with respect to the components of the SLO process (objective statement, performance criteria, rubric, performance-based tasks), but it also makes the resulting analyses something of a best case scenario[5].

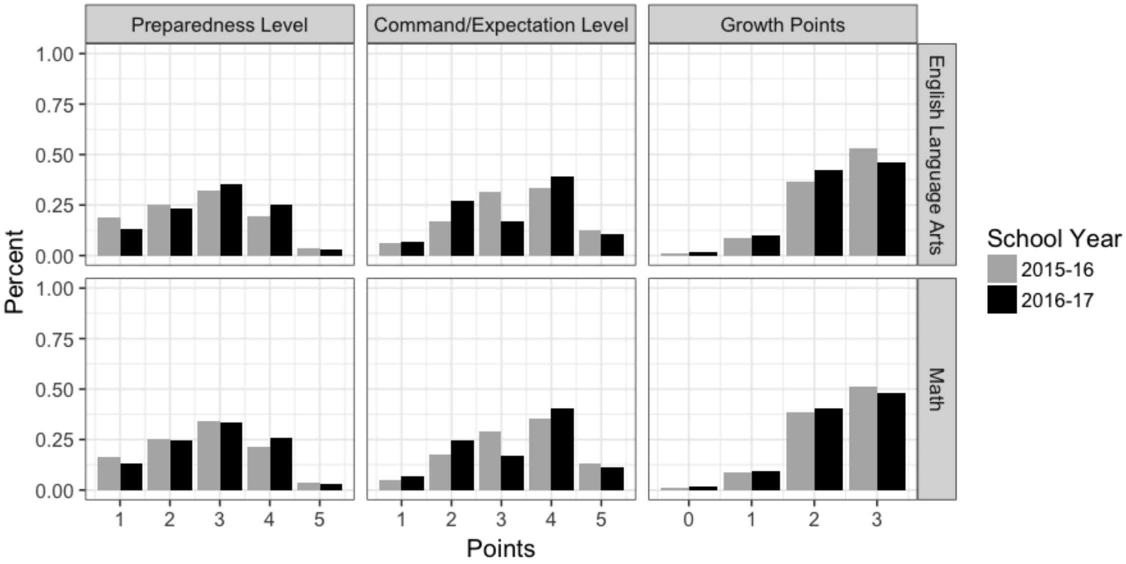
[5] The teachers in each subject domain were generally distinct in 2015-16 because in this year teachers were only required to submit a single SLO, though some did submit two. In 2015-16 there were a total of 1,974 unique teachers and 313 who submitted an SLO for both ELA and math. In 2016-2017 (when teachers were required to submit two SLOs) there were a total of 2,025 unique teachers and 735 who submitted in both subject domains. The overlap across subject domain samples for students was 9,067 and 19,030 for 2015-16 and 2016-17 respectively. Importantly however, for both years when conditioning on subject domain, sample sizes represent unique teachers and students. In a relatively small number of cases, the same student was present for the SLO of more than one teacher within the math and ELA subject domains. In these instances, to ensure that the same student was only present once in our analyses, we randomly assigned the student to a single teacher.

Table 1. Analytic Populations for Analyses

	All SLOs		District Templates			
			ELA SLOs		Math SLOs	
	2015-16	2016-17	2015-16	2016-17	2015-16	2016-17
Students	56,295	61,023	25,115	27,463	18,238	28,671
Teachers	4,267	4,383	1,415	1,444	872	1,316
Schools	135	140	126	133	120	132

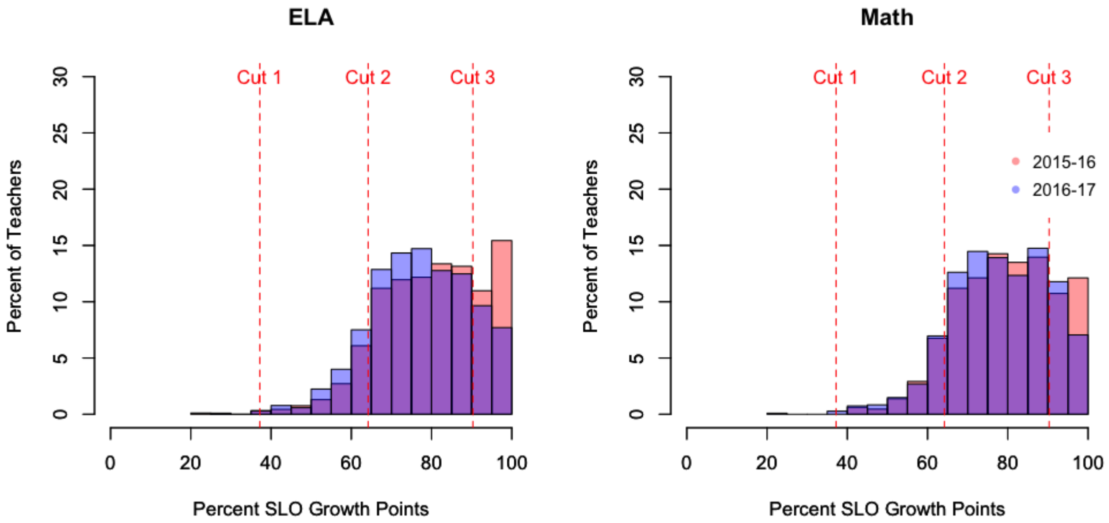
Figure 4 presents frequency distributions of SLO preparedness, end-of-year command and growth points across all students in both academic school years. Interestingly, these distributions are almost identical by year and by subject domain. In both years, the average DPS student was classified as “somewhat prepared” for his or her SLO at the start of the school year (mean between 2.6 and 2.7), and as having “moderate command” of the SLO by the end of the instructional period (mean between 3.2 and 3.3). Most students received 2 or 3 growth points out of a maximum possible of 3 points (after applying the scoring rules shown in Figure 3). This suggests that, according to their teachers, the vast majority of DPS students showed evidence of “moderate” to “high” growth.

Figure 4. Distributions of SLO outcomes for ELA and Math students by Year



The SLO variable of interest at the teacher level is the percent of SLO points earned by each teacher. On the basis of their percent of SLO points earned (which we will refer to henceforth as “SLO.Growth”) and cut scores set by DPS leadership, teachers fell into one of four effectiveness categories. For example, out of 1,415 teachers with ELA SLOs in 2015-16, 0.3% were rated “ineffective,” 8.6% of teachers were rated “approaching,” 60.5% were rated “effective,” and 30.6% were rated “distinguished.” One year later in 2016-17, out of 1,444 teachers with ELA SLOs, 0.3% were rated “ineffective,” 15.4% were rated “approaching,” 72.5% were rated “effective,” and 11.7% were rated “distinguished.” Similar changes occurred for Math. Figure 5 displays the distribution of this variable by SLO subject and year with the cut points for LEAP effectiveness categories from 2015-16 superimposed.

Figure 5. Distribution of Percent of SLO Growth Points Earned by a Teacher by Subject and Year.



The distribution of the SLO Growth variable has a strong skew toward high values. To put this in perspective, one third of teachers with ELA SLOs in 2015-16 (486 out of 1,415) and about a quarter in 2016-17 (315 out of 1,444) earned 90% or more of possible SLO growth points. At the extremes, 14% and 6% in each year earned 100% of possible SLO growth points. This is indicative of a ceiling effect on teacher growth scores, although the effect appears to have diminished to a considerable extent from the first to second year of full SLO implementation. The skew towards high values 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 by the way teachers classified students with respect to preparedness and end of year command. We explore this latter hypothesis in the next section.

Validity of SLO Preparedness and Command Classifications

Preparedness Classifications

To examine the validity of teachers' SLO preparedness classifications, we use as an external indicator students' ELA and Math PARCC scores from the previous school year (either the spring of 2015, or the spring of 2016) as a criterion to characterize the consistency of teachers' decisions. An important limitation of this analysis is that it can only be performed for students with ELA and/or Math SLO classifications who have PARCC scores from the preceding year. For example, using 2015-16 data, this restriction reduces our student sample from 25,115 to 9,760 in ELA, and from 18,238 to 8,642 in Math.

Table 2 shows the cross-tabulation of these two student classifications using 2015-16 data (we created and examined a similar cross-tabulation with 2016-17 data). There are a number of different ways that one could characterize preparedness classifications as consistent with PARCC test score performance. The strictest rule would be a one to one agreement between PARCC performance levels and SLO preparedness levels such that only cells along the main diagonal of each cross-tabulation are considered consistent. Using this criterion with 2015-16 data, we would find that 39.9% of ELA and 41.5% of Math SLO preparedness classifications were consistent. Using 2016-17 data, these values increase to 47.3% and 53.3%, respectively, likely because prior grade PARCC scores were available for teachers to consult when making preparedness ratings in 2016-17, but not in 2015-16.

Table 2. SLO Preparedness Levels by PARCC Performance Levels: 2015-16 Data

PARCC Performance Level Spring 2015	SLO Preparedness Level Fall 2015					Totals
	Sig. Up. [1]	Up. [2]	S. Prep. [3]	Prepared [4]	Ahead [5]	
English Language Arts						
Did not yet meet expectations [1]	1,086	680	378	45	2	2,191
Partially Met Expectations [2]	477	815	761	160	3	2,216
Approached Expectations [3]	211	603	985	486	36	2,321
Met Expectations [4]	73	321	984	912	190	2,480
Exceeded Expectations [5]	6	34	171	245	96	552
Total	1,853	2,453	3,279	1,848	327	9,760
Mathematics						
Did not yet meet expectations [1]	686	534	230	27	1	1,478
Partially Met Expectations [2]	554	982	818	134	4	2,492
Approached Expectations [3]	184	626	965	481	20	2,276
Met Expectations [4]	59	205	762	853	184	2,063
Exceeded Expectations [5]	1	3	73	157	99	333
Total	1,484	2,350	2,848	1,652	308	8,642

Notes: Cells with gray shading are considered “consistent SLO classifications” and those without shading are considered to be “inconsistent.”

However, there is no particular reason to require or expect a one to one relationship between PARCC performance levels and SLO preparedness levels. An alternative approach is to regard the PARCC performance levels of 3 (“approached expectations”) and 4 (“met expectations”) as a key dividing line. 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 gray in Table 2), and a preparedness level of 1, 2 or 3 to be an inconsistent classification (cells that are not shaded). For a student in PARCC performance level of 3 or lower, so long as the student’s preparedness level is also 3 or lower, we consider it a consistent classification (cells shaded gray), while a preparedness level of 4 or 5 is considered an inconsistent classification (cells that are not shaded). In summary, using this dichotomous classification criterion, we would find that 76.2% of ELA and 79.5% of Math SLO preparedness classifications were consistent in 2015-16, with these rates increasing to 80.6% and 84.5% in 2016-17.

A closer look at Table 2 suggests 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 3,032 and 2,396 students who either met or exceeded expectations on the PARCC tests for ELA and Math respectively (performance level 4 or 5). Only 48% and 54% of these students were classified consistently—in other words about 52% and 46% of these students were classified as significantly underprepared, underprepared or only somewhat prepared. In contrast, consider the 6,728 and 6,246 students who were classified as approaching expectations or lower on the PARCC tests for ELA and Math (performance level 1, 2 or 3). Only about 10 to 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. The top half of Table 3 summarizes these different types of SLO preparedness classification rates by subject domain and year.

Table 3. Summary of Consistency with PARCC Performance for SLO Preparedness and End of Year Command Classifications

	ELA		Math	
	2015-16	2016-17	2015-16	2016-17
SLO Year Preparedness				
Strict Consistency %	39.9	47.3	41.5	53.3
Dichotomous Consistency %	76.2	80.6	79.5	84.5
Underestimation %	52.4	39.1	46.0	35.8
Overestimation %	10.9	8.8	10.7	7.0
SLO End of Year Command				
Strict Consistency %	40.4	39.6	42.7	39
Dichotomous Consistency	76.6	78.7	79.1	80.5
Underestimation %	25.5	20.7	18.0	11.8
Overestimation %	22.4	21.7	22.3	23.2

It is important to appreciate that an inconsistent classification as defined here (i.e., when a student's SLO preparedness appears to be underestimated or overestimated), is not necessarily an *inaccurate* classification. It is entirely possible that a student who performed at a high level on the prior grade PARCC test might not show evidence of being "prepared" for an SLO at the beginning of the next school 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 be attributable to summer learning loss.

End-of-Year Command Classifications

Here we follow an approach that is somewhat parallel to the one above, only this time we compare SLO end-of-year command classifications of students (as of spring 2016 or spring 2017) to the performance of these students on the same subject PARCC test (also taken in the spring of 2016 or the spring of 2017). The bottom half of Table 3 summarizes these results. With respect to the overall consistency of end-of-year classifications, the results are almost identical to what was found for preparedness classifications. That is, taking a strict approach to consistent classification (agreement along the main diagonal), only about 40 and 43% of student were consistently classified for ELA and math SLOs respectively. Using a dichotomous classification approach, the values jump to 77 and 79%. These values stayed about the same in 2016-17.

If teachers have an incentive to inflate their growth scores by giving higher end-of-year command ratings than their students merit, a place to look for evidence of this would be the students who were in the bottom three performance levels of the PARCC ELA and Math tests. With both 2015-16 and 2016-17 data we find that in both subject domains about 22% of students appear to have had their end-of-year rating overestimated. These are students who did not "meet expectations" on the PARCC tests taken in March but were classified as "strong" or "distinguished" on their SLO in May. This is primarily driven by students placed into the strong command categories—fewer than 1% of students in the bottom three PARCC performance levels were placed into the "distinguished" SLO category.

Is the Preparedness of Certain Kinds of Students Underestimated?

Next, we examine whether certain kinds of students were more likely than others to have their preparedness underestimated. We first do this by conducting ordered logit (i.e., ordinal) regressions (e.g., Long, 1997) with students as the units of analysis. The dependent variable in these regressions is the preparedness level of the student from 1 to 5. The independent variables consist of within-grade standardized PARCC scale scores from the previous year and student-specific demographic dummy variables that take on values of 1 if a student is female (“Female”), nonwhite (“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. We run this regression four times, once for each SLO subject domain (math and ELA), and once for each academic school year (2015-16 and 2017-18). The respective student sample sizes for ELA and Math SLOs were 9,749 and 8,630 in 2015-16; 10,463 and 12,017 in 2016-17. Of primary interest is whether the demographic covariates are predictive of SLO preparedness levels, even after controlling for prior year PARCC scale scores.

Table 4. Student-level Ordered Logit Regressions (Odds Ratios)

	Dependent variable: Ordinal Categories 1-5							
	SLO Preparedness				SLO End-of-Year Command			
	2015-16		2016-17		2015-16		2016-17	
	ELA	Math	ELA	Math	ELA	Math	ELA	Math
PARCC.SS	3.32***	3.25***	5.44***	6.89***	4.78***	6.20***	3.60***	4.52***
Female	1.07	1.04	1.00	0.94	1.13***	1.11**	1.17***	1.15***
NonWhite	0.92	0.78***	0.99	0.89*	0.96	0.96	0.90*	0.90*
FRL	0.87*	0.76***	0.84**	0.83***	0.93	1.03	0.87	0.80
ELL	1.01	0.86**	1.11*	0.92*	0.91*	1.05	1.09*	1.08*
IEP	0.33***	0.40***	0.34***	0.42***	0.43***	0.53***	0.79***	0.72***
GT	1.30***	1.52***	1.07	0.94	1.41***	1.35***	1.58***	1.44***
Students	9,749	8,630	10,463	12,017	12,146	9,802	13,303	15,956

Notes: *p<0.1; **p<0.05; ***p<0.01. The results in each column are instances of an ordered logit regression model with four thresholds identified by fixing the constant at 0. Threshold estimates are not included here but are available upon request.

Columns 2 through 5 of Table 4 presents the results from these ordinal regressions, with coefficients expressed in an odds ratio metric. In both subjects across years, as one would expect, the odds of a student being in a higher preparedness category or categories relative to the lower category or categories increases dramatically for a 1 SD increase in prior PARCC scores. Here we see again that the coefficients for the odds ratios increase significantly from 2015-16 to 2016-17, consistent with the results shown earlier, and the availability of prior grade PARCC scores in 2016-17, but not in 2015-16. We also see that students with an IEP are much more likely to be in a lower preparedness category, while GT students have greater odds of being in a higher preparedness category. Neither of these results is surprising, since each variable represents sources of information teachers would be expected to consult when making a classification along with prior year test performance.

Of concern is the significant finding across subjects and years that poorer students (those eligible for free and reduced price lunches), have a slightly higher odds of being in lower preparedness categories. While the magnitude of this association is small, the fact that it is statistically significant, even after controlling for prior PARCC performance is worrisome, because this represents information that teachers should not be using to judge SLO preparedness. Similarly, there is some evidence that the odds of landing in a higher preparedness category is related to a student's status as an English Language Learner and for NonWhite students on Math SLOs (though again, the magnitudes tend to be small, and for these variables the direction and statistical significance are not always consistent).

The ordinal regression results presented in Table 4 make a proportional odds or parallel regression assumption. That is, it is assumed that the odds of shifting from one category to the next is the same across all categories. A closer inspection of this assumption suggests that it may not hold for the covariates NonWhite and FRL for the two highest SLO preparedness categories. To dig into this more deeply, we restricted our focus to the subset of students who scored in performance levels 4 or 5 on the PARCC ELA and Math tests. We then specify logistic regressions where 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. This logistic regression includes the same set of demographic variables as covariates that were in the ordinal regression described above, but does not include prior year PARCC scale scores since prior year PARCC performance levels are being used directly to both restrict the sample and define the outcome variable. In this approach, we again look for evidence that students with certain demographic characteristics have their preparedness underestimated, but we do so by focusing on a subset of students (those who met expectations on PARCC the previous spring), and a specific preparedness threshold (the three lowest preparedness levels vs. the two highest).

In taking this approach we find some evidence of a larger magnitude of underestimation for students who are FRL eligible and/or nonwhite. Using 2016-17 SLO data for ELA as an example, a nonwhite male who is eligible for free and reduced lunch, a native English speaker, with no IEP and not identified as Gifted and Talented had a probability of 51.7% of being classified as Somewhat Underprepared, Underprepared, or Significantly Underprepared—even though the student had scored in one of the top two PARCC ELA performance levels about 4-5 months earlier. In contrast, if the student differed only by being white and not FRL eligible, the probability of a low preparedness classification would decrease by almost 17 points to 35 %. The marginal change in probability is similarly large for math SLOs, where the race/ethnicity and FRL indicators increase the probability of underestimation about 13 points (from 28.5% to 41.6%). These marginal changes in probability within SLO subject domain were fairly similar for other reference students with different combinations of the demographic covariates.

Is the End-of-Year Command of Certain Kinds of Students Overestimated?

We performed a parallel set of analyses with the end-of-year SLO command classifications as the outcome variable of interest. Columns 6 through 9 of Table 4 present the results from the ordinal regressions, with coefficients expressed in an odds ratio metric. Once again, we similar relationships between PARCC scores, IEP and GT indicators, and classification in higher end of year command category as was found for preparedness classifications. However, we see no association with race, poverty of English Learner status. Instead, females tend to be slightly more likely than males to be in a higher end of year command category.

Relationship of Teacher SLO Growth Points to Other Teacher-Level Variables

In this section, we shift from examining the validity of student-level preparedness and end-of-year command classifications to the validity and reliability of using SLO.Growth scores as a measure of teacher effectiveness. We do so by adopting a framework that has been used in the past 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 (e.g., Ehlert, Koedel, Parson & Podgursky, 2014). 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 student achievement conditional on students' starting points designated during the fall. 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 preparedness and mastery are determined with some degree of subjectivity by teachers. Because value-added models take pre-existing differences in student achievement (and often other variables) into account, in theory at least, 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 growth measure based on SLOs appears to "level the playing field" in a manner analogous to a value-added model when contrasted to the use of a teacher-level status measure based on SLOs.

We examine teachers with ELA and Math SLOs separately in this analysis, and also impose a restriction that each teacher must have at least 15 students. This reduces our teacher samples shown in Table 1 by about 29 to 35% in 2015-16, and 18 to 25% in 2016-17. In what follows, we contrast two different teacher-level SLO outcome variables. The first, "SLO.Status" is the average of a students' end-of-year SLO mastery classifications (the average of ordinal values on a scale from 1 to 5). The second, "SLO.Growth" represents the percent of possible growth points earned by a teacher (on a scale from 0% to 100%). Table 5 shows correlations between these two variables as well as between other available and relevant teacher-level variables for two different teachers samples in 2015-16: the 920 teachers that submitted ELA SLOs (upper right triangle above the main diagonal) and the 652 that submitted Math SLOs (lower left triangle below the main diagonal). The main diagonal provides an estimate of the SD of each variable, computed as the average of the ELA and Math teacher samples. A variable of particular interest is each teacher's professional practice rating ("PP.Pts"). This score represents a combination of classroom observation ratings, a professionalism rating, and student perception survey ratings. The remaining covariates are all created by computing teacher-level means for the students associated with each teacher. These include the mean subject-specific PARCC scale scores at the teacher level for two different years ("PARCC.15" and "PARCC.16"), the mean of current year student growth percentiles ("MGP.16") and all the aggregate demographic variables used in the student-level logistic regressions previously presented. Because the PARCC tests are not on the same scale across grades, all scores were standardized across the full population of students in the district by grade and subject for a given year. We only present this correlation matrix for 2015-16 data; the results for 2016-17 data were very similar (within about .05 on a scale from 0 to 1).

Table 5. Relationships among teacher-level variables (2015-16).

	SLO. Status	SLO. Growth	PP.Pts	MSS.15	MSS.16	MGP.16	%Fem	%Non White	%FRL	%ELL	%IEP	%GT
SLO.Status	(0.60)	0.49	0.29	0.54	0.56	0.22	-0.01	-0.48	-0.49	-0.45	-0.23	0.09
SLO.Growth	0.46	(0.13)	0.21	0.06	0.11	0.13	0.04	-0.04	-0.04	-0.11	-0.08	-0.07
PP.Pts	0.33	0.25	(4.49)	0.24	0.30	0.26	0.06	-0.28	-0.30	-0.20	-0.15	0.14
PARCC.15	0.64	0.07	0.28	(0.62)	0.93	0.31	0.14	-0.81	-0.85	-0.66	-0.37	0.72
PARCC.16	0.63	0.08	0.36	0.90	(0.59)	0.62	0.12	-0.80	-0.83	-0.60	-0.25	0.61
MGP.16	0.29	0.12	0.34	0.27	0.60	(13.26)	0.10	-0.31	-0.30	-0.22	-0.11	0.29
%Fem	0.00	0.02	0.04	0.02	0.01	-0.03	(0.10)	-0.01	-0.02	-0.01	-0.14	0.12
%NonWhite	-0.53	-0.04	-0.31	-0.79	-0.81	-0.30	0.03	(0.30)	0.94	0.65	0.18	-0.31
%FRL	-0.54	-0.04	-0.30	-0.81	-0.84	-0.33	0.01	0.95	(0.33)	0.65	0.20	-0.35
%ELL	-0.46	-0.10	-0.25	-0.58	-0.61	-0.27	-0.02	0.68	0.67	(0.33)	0.07	-0.12
%IEP	-0.17	0.04	-0.05	-0.36	-0.24	-0.14	-0.08	0.16	0.19	0.08	(0.08)	-0.17
%GT	0.14	-0.07	0.06	0.72	0.63	0.27	0.07	-0.27	-0.32	-0.11	-0.17	(0.17)

Note: Correlations for ELA teacher sample in the upper triangle, correlations for math teacher sample in the lower triangle. Correlations were computed using pairwise complete observations. The values in parentheses along the main diagonal represent the average standard deviation of each variable for the two samples.

Notice in Table 5 that SLO.Status tends to be more strongly correlated with aggregated demographic and achievement variables than SLO.Growth. In particular, while the teacher-level mean of prior year PARCC scale scores in ELA and Math have correlations of about .54 and .64 with the teacher-level mean of SLO.Status, the two variables are essentially uncorrelated with SLO.Growth.

We explore these relationships more formally by regressing (with teachers as the unit of analysis) aggregate ELA and Math 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 teacher scores for professional practice[6].

$$\begin{aligned} Tch.SLO.Outcome \\ = \beta_0 + \beta_1 PP.Pts + \beta_2 PARCC.lagged + \beta_3 \%Female + \beta_4 \%FRL + \beta_5 \%ELL \\ + \beta_6 \%IEP + \beta_7 \%GT + \beta_8 Middle + \beta_9 High + \beta_{10} Other\ School + \varepsilon. \end{aligned}$$

In the model above, $Tch.SLO.Outcome$ is a teacher's average of student's end-of-year SLO mastery classification (SLO.Status) or a teacher's percent of SLO points earned (SLO.Growth). The variable $PP.Pts$ represents the professional practice points a teacher earned in the school year in consideration, and $PARCC.lagged$ represents the mean standardized student PARCC scores from either spring 2015 (for 2015-16 SLO results) or spring 2016 (for 2016-17 SLO results). The variables $Middle$, $High$ and *Other School*[7] capture school level differences with Elementary as the omitted reference category.

[6] As shown in Table 5, once aggregated to the teacher-level, the variables representing the percentage of nonwhite students (%NonWhite) and percentage of students eligible for free and reduced price lunches (%FRL) have a correlation of about 0.94. To avoid problems with a severe form of multicollinearity we only retain %FRL as a covariate in our regression models.

[7] "Other School" represents schools with the atypical grade configurations of K-8, 6-12, or K-12.

Finally, $\%Female$, $\%FRL$, $\%ELL$, $\%IEP$, and $\%GT$ 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. We run separate regressions for each year, subject domain and SLO outcome. To facilitate comparisons across models, we standardize all regression variables with the exception of our school level indicators for each subject domain and year combination. The estimated coefficients can be found in Table 6 and are interpretable as the number of SDs the dependent variable of interest would be predicted to increase for a 1 SD increase in the independent variable of interest, holding other covariates constant.

Table 6. Regression Models for Teachers' SLO Outcomes

	ELA 2015-16		Math 2015-16		ELA 2016-17		Math 2016-17	
	SLO. Status	SLO. Growth	SLO. Status	SLO. Growth	SLO. Status	SLO. Growth	SLO. Status	SLO. Growth
PP.Pts	0.15^{**}	0.20^{***}	0.15^{**}	0.20^{***}	0.14^{***}	0.33^{***}	0.14^{***}	0.27^{***}
PARCC.lagged	0.48^{***}	0.17	0.35^{***}	0.03	0.28^{**}	-0.05	0.42^{***}	0.07
%Fem	-0.03	0.02	0.03	-0.05	-0.05	0.03	0.02	0.08
%FRL	-0.04	0.10	-0.16	0.11	-0.21^{*}	-0.002	-0.21^{**}	-0.05
%ELL	-0.03	0.02	0.04	0.15	0.03	-0.005	-0.003	0.10
%IEP	-0.08	0.01	-0.04	0.002	-0.03	0.01	-0.03	0.003
%GT	-0.05	-0.08	0.15^{*}	0.20^{*}	0.23^{***}	-0.001	0.13^{**}	0.01
Middle	0.04	-0.23	-0.19	-0.06	-0.31^{**}	-0.14	-0.37^{***}	0.08
High	-0.37^{**}	-0.06	-0.38^{**}	-0.33^{*}	-0.54^{***}	-0.14	-0.52^{***}	0.12
Other School	-0.35^{*}	-0.16	-0.04	0.02	-0.17	-0.43^{**}	-0.05	-0.16
Constant	0.11	0.07	0.14^{*}	0.09	0.17^{**}	0.11	0.14^{**}	-0.004
Observations	343	343	301	301	380	380	434	434
R^2	0.35	0.05	0.46	0.10	0.47	0.12	0.59	0.10

Note: * p<0.05; ** p<0.01; *** p<0.001

Two patterns in these results stand out. First, covariates such as %GT, %IEP, %FRL and school-level indicators, which frequently have a significant partial association with SLO.Status for a given subject domain and year, seldom have a significant partial association with SLO.Growth. Second, the only covariate that retains a significant association with both SLO.Status and SLO.Growth outcomes is a teacher's professional practice score (PP.Pts). In fact, though the differences tend to be relatively small, the regression coefficient for the PP.Pts covariate is always larger in magnitude when regressed on the SLO.Growth outcome relative to the SLO.Status outcome. For example, for 2015-16 data, a 1 SD increase in a teacher's professional practice score is associated with about a .15 SD increase on SLO.Status scores. In contrast, a 1 SD increase is associated with a about a .20 SD increase on SLO.Growth scores.

Comparisons with Mean Student Growth Percentiles

In principle, both an SGP and SLO are supposed to convey something about student growth that, when aggregated to a teacher, can be used to discern that some teachers have been more effective than others in their academic instruction. An important question is whether a growth indicator based on SLOs (i.e., SLO.Growth) is an adequate substitute for a measure based on SGPs (i.e., MGP). To address this question, we examine the slightly smaller subset of teachers from the regression analysis above for whom MGP scores are also available. Figure 6 displays the scatterplots of SLO.Growth and MGP by subject domain and year. Although both SLO.Growth and MGP are intended to be teacher-level indicators of student growth, with correlations ranging between 0.12 and 0.28, they tend to not be much more strongly correlated with one another than they are with professional practice ratings, even though the latter are intended to capture a different dimension of teacher effectiveness[8].

Figure 6. Relationship Between SLO Points and Mean of Student Growth Percentiles

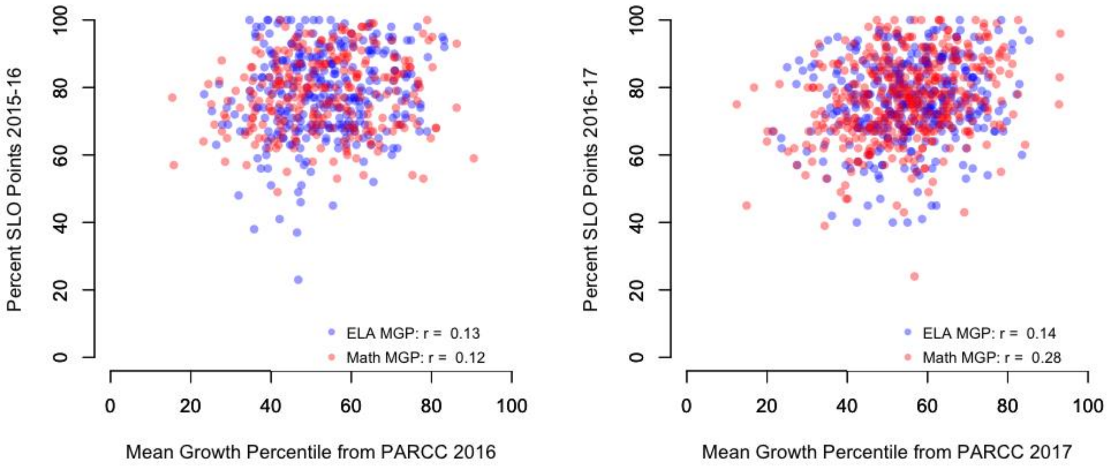


Table 7 replicates the format of Table 6, but adds a column in which we replace the SLO-based aggregate outcome (SLO.Status or SLO.Growth) with an SGP-based aggregate outcome (i.e., a teacher's MGP). There are again two interesting and interpretable findings[9]. First, the R^2 for regressions with MGP as the outcome are always

[8] In both subject domains, the correlation between SLO.Growth and MGP increased from 2015-16 to 2016-17. Again, this finding may well be attributable to the impact of having prior year PARCC scores available when making preparedness ratings in 2016-17, but not having them available in 2015-16.

[9] A puzzling finding is the statistically significant regression coefficients of −.38 and −.31 for the covariates %FRL and PARCC.lagged when math MGP is the outcome using 2015-16 data. After ruling out coding errors as an explanation and performing a variety of regression diagnostics, our conclusion is that this difference relative to the results in the other three models is most likely being caused by some interaction between sample size (this year and subject domain had the smallest teacher sample) and multicollinearity. That is, all four of the regressions shown in Table 7 have the same issues with multicollinearity among the independent variables induced by the aggregation of student-level variables. But only in this one case does it manifest itself with a shift in the magnitude and sign of a regression coefficient that does not cohere with theory or findings from previous studies.

higher than the R^2 for regressions with SLO.Growth as the outcome. More specifically, the covariate %GT maintains a significant partial association with MGP for each subject domain and year contribution; a similar association is not evident for the SLO.Growth outcome. Second, the professional practice score has about the same partial association with the MGP outcome (ranging from a low of .21 to a high of .30) as it does for the SLO.Growth outcome (ranging from a low of .20 to a high of .34).

If one criterion for the validity of a classroom growth indicator for use in teacher evaluation is that it should "level the playing" field such that teachers are not rewarded or penalized for characteristics of students that are outside of their control, then an argument could be made based on these results that the SLO.Growth indicator fulfills this criterion as well or better than the MGP indicator. On the other hand, one might argue that for a growth indicator to be valid that it *should* retain at least a small positive correlation with a variable such as %GT, under the premise that students flagged with gifted and talented status have typically demonstrated more growth each year than students not flagged with this status. However, while the results are open to slightly different interpretations, the bottom line is the regression relationships tend to be remarkably similar whether one uses SLO.Growth or MGP as the teacher-level dependent variable. This suggests the need for caution when using this as the basis for a validity argument, and we return to discuss the implications of these comparisons in the final section of the paper.

Table 7. Regression Models for Teachers with MGPs

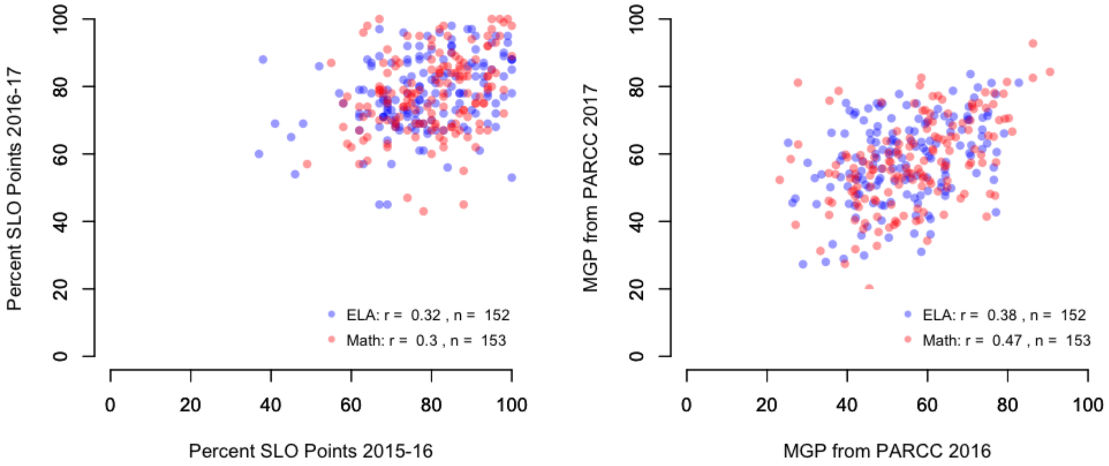
	ELA 2015-16			Math 2015-16			ELA 2016-17			Math 2016-17		
	SLO. Status	SLO. Growth	MGP	SLO. Status	SLO. Growth	MGP	SLO. Status	SLO. Growth	MGP	SLO. Status	SLO. Growth	MGP
PP.Pts	0.10*	0.20**	0.21***	0.11*	0.21**	0.29***	0.13**	0.34***	0.30***	0.14***	0.27***	0.29***
PARCC.lagged	0.38**	0.16	0.01	0.41***	0.06	-0.31*	0.29**	-0.01	0.03	0.45***	0.05	-0.02
%Fem	-0.02	0.01	0.08	0.02	-0.07	-0.03	-0.03	0.04	-0.05	0.01	0.07	-0.04
%FRL	-0.08	0.14	-0.11	-0.17	0.12	-0.38*	-0.22*	0.05	-0.09	-0.17*	-0.06	-0.14
%ELL	-0.10	-0.02	-0.05	0.01	0.16	0.02	0.05	-0.03	0.01	-0.02	0.10	0.08
%IEP	-0.09	0.02	0.02	-0.03	-0.01	-0.12*	-0.03	0.01	-0.07	-0.03	-0.02	0.03
%GT	0.06	-0.03	0.19*	0.09	0.17	0.22*	0.24***	0.01	0.18*	0.13**	0.01	0.16*
Middle	0.06	-0.22	-0.29	-0.15	-0.03	0.04	-0.31**	-0.14	-0.01	-0.37***	0.07	-0.32*
High	-0.61**	0.02	0.04	-0.17	0.002	0.17	-0.57***	-0.50*	0.10	-0.42**	0.13	-0.18
Other School	-0.33*	-0.15	-0.08	-0.01	0.02	0.13	-0.17	-0.42**	0.09	-0.05	-0.16	-0.22
Constant	0.09	0.06	0.06	0.05	0.004	-0.05	0.14*	0.14	-0.02	0.11*	0.004	0.11
Observations	298	298	298	249	249	249	347	347	347	407	407	407
R^2	0.40	0.05	0.17	0.48	0.08	0.23	0.50	0.14	0.22	0.61	0.10	0.15

Note: * p<0.05; ** p<0.01; *** p<0.001

Stability

Beyond concerns about their validity as measures of teacher effectiveness, a key drawback to any growth-based statistic is that measures of growth tend to be more volatile than measures of status. In the literature on value-added models, the year to year correlation of teacher growth statistics has been found to be weak to moderate, ranging from about 0.2 to 0.6 (Goldhaber & Hansen, 2008; McCaffrey et al., 2009). Kane & Staiger (2002; 2008) and McCaffrey et al. (2009) have argued that under certain assumptions, such intertemporal correlations can be interpreted as an estimate of stability, in which case any intertemporal correlation less than 0.5 would imply that more than half of the variability in value-added can be explained by chance factors unrelated to characteristics of a teacher that persist over time.

Figure 7. Intertemporal Stability of Teacher-Level Growth: SLO vs. SGP



The scatterplots in Figure 7 depict the relative stability of teacher-level SLO and SGP growth statistics when focusing on the subset of teachers for whom both statistics were available across two years of data. The intertemporal correlations tend to be small to moderate, ranging from about .30 to .47. The stability of growth based on SLOs is lower than the stability of growth based on SGPs. For ELA the stability is just slightly higher for MGPs (*r* = .38 instead of *r* = .32), but for Math it is considerably higher (*r* = .47 instead of *r* = .30). Nonetheless, the lower correlation for SLOs is well within the range of what is typically observed for most growth-based statistics (McCaffrey et al., 2009).

These results need to be interpreted with some caution since to the extent that each of these teacher-level statistics is biased, the correlations may be overstated to an unknown degree. For example, MGPs make no attempt to disentangle influences on student achievement beyond those captured by a student's individual prior test achievement profile. Assume that a student's classroom peers have a distinct effect on student achievement. If some teachers tend to always work with students in classrooms with positive peer effects, and others tend to work with students in classrooms with negative peer effects, then the former teachers will always be more likely to have an MGP that is above average in any given year, and the latter teachers will always be more likely to have an MGP that is below average in any given year. This bias could manifest itself through inflated intertemporal correlations. The potential for bias is just as great (if not greater) for measures of growth based on SLOs. In this context, one might imagine two different sources of bias, one due to omitted variables associated with both classroom contexts and student achievement, and another due to the fact that teachers are the ones classifying the preparedness and end of year command of their own students. If certain teachers are more likely to underestimate the preparedness of their students every year and others are more likely to overestimate it, this could also inflate the observed stability of SLO growth indicators.

Discussion

In the way that they are implemented in Denver Public Schools, SLOs are premised on a dual purpose use for classroom assessment. That is, teachers are expected to use classroom assessments that they choose, administer and score so that they can track and facilitate student learning over the course of the year. At the same time, the result of this process also includes numeric ratings that get aggregated into a teacher-level growth indicator, which in turn figures prominently in teachers' annual performance evaluations. The *Standards for Educational and Psychological Testing* offer a general warning when it comes to the use of educational assessments to meet multiple purposes: "Most educational tests will serve one purpose better than others; and the more purposes an educational test is purported to serve, the less likely it is to serve any of those purposes effectively." (AERA/APA/NCME, 2014, p. 206)

In the present study we do not formally evaluate the claim that SLOs are a successful vehicle for formative assessment. However, the results from focus group interviews and surveys that we have conducted (but do not present here) do not paint an encouraging picture in this regard (Authors, blinded). For example, in one survey administered in early 2016, only about 60% of a sample of 1,712 DPS teachers agreed or strongly agreed with the prompt "SLOs provide me valuable information about what my students know and can do." While these findings have somewhat limited generalizability because of self-selection among survey (and focus group) respondents and our inability to link these respondents to the data considered in this paper, they suggest at a minimum that many DPS teachers have not bought into the dual purpose use of SLOs. Instead, they tend to view SLOs almost exclusively as something required of them to meet district accountability demands.

What we have looked for in this study is evidence, through available data over a two year time period, that student and teacher-level growth statistics based on the SLO performances of students are being distorted as one might predict from Campbell's Law. We have also compared the use of an SLO-based growth statistic for teacher evaluation to the use of an SGP–based growth statistic. When using PARCC test performance as an external criterion, we find evidence to suggest that teachers may have a tendency to underestimate some students' SLO preparedness and/or overestimate some students' end-of-year level of SLO command, two practices that would lead to an inflation of SLO growth points. The underestimation of preparedness appears to be a bigger concern than the overestimation of end-of-year command. Using 2015-16 data restricted to the population of students that were in one of PARCC's two highest performance levels in the Spring of 2015, we found that roughly 50% of these students had SLO preparedness classifications that seemed too low. In contrast, when using the same data and restricting to the population of students that scored in one of PARCC's three lowest performance levels near the end of the school year, we found that only about 10 to 11% of students had SLO end of year command classifications that seemed too high.

The general finding that many students had preparedness classifications that were lower than what was suggested by their PARCC performance has at least two possible explanations. One explanation is that many teachers had the (mistaken) impression that SLO preparedness and end-of-year 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, more consistent with Campbell's Law, would be that at least some teachers, aware that student growth would count toward their LEAP evaluation ratings, responded to this incentive by placing students into lower preparedness levels to maximize their growth ratings. These two explanations are not mutually exclusive.

A more troubling finding is that certain characteristics of students—in particular gender, race/ethnicity, and poverty status—are often predictive of having preparedness underestimated or end-of-year command overestimated. Although the magnitude of this underestimation or overestimation tended to be small overall, it was not nonexistent, and it appears to have an interaction with race and poverty. Holding constant prior academic achievement in the form of PARCC performance, students in poverty are still significantly more likely to have their preparedness underestimated relative to students not in poverty.

SLO growth points are typically aggregated to create a teacher-level growth measure. For teachers that have both SLO-based growth measures and SGP-based growth measures, the correlation between the two is weak. Somewhat surprisingly, with respect to its relationship to other teacher-level variables and with respect to its intertemporal stability, SLO.Growth is difficult to distinguish from an MGP. Both SLO.Growth and MGPs have the same low (but statistically significant) correlation with professional practice ratings, and this correlation (not corrected for measurement error) is of about the same magnitude as that found between value-added estimates and teacher professional practice ratings in at least one other prominent empirical context (Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger; 2012).

If an advantage of measures of growth over status is that they are less correlated (or uncorrelated) with aggregate measures of student status, than one could argue that SLO.Growth displays this advantage as well or better than MGPs. Both SLO.Growth and MGPs have a weak to moderate degree of intertemporal stability. In general, the finding of low intertemporal stability for teacher or school-level measures of growth has been a key to the argument that such measures should be employed with great caution for high-stakes purposes (Morganstein & Wasserstein, 2014). This argument certainly applies in this context as well, and it is was a greater concern for SLO.Growth than it was for MGP, especially in Math.

There are limitations to the generalizability of findings in this study. In our analyses we generally focus on the non-random sample of DPS teachers who submitted district-created SLOs in Math and ELA, and who also teach students for whom large-scale assessment results were available. Recall that one of the biggest motivations for SLOs is that they provide growth evidence for teachers who teach in subjects for which state-administered large-scale assessments are *not* available. If the SLOs for these teachers tend to be of lesser quality or involve more distorted practices than those analyzed in this study, our findings may well present an overly optimistic evaluation.

Another limitation is that we cannot claim that the use of SLOs for high-stakes has *caused* a distortion in SLO scores. It is entirely possible that the evidence we have found with respect to under and overestimation would be evident even in the absence of the use of SLOs for accountability. It would be interesting to replicate our approach in a comparable school district or an entire state that uses SLOs, but only for formative purposes.

To what extent do the analyses in the second half of this paper suggest that SLO.Growth can be validly used as not just an alternative to MGPs for accountability decisions, but as a replacement? Although some of our findings might imply support to this notion, we would caution readers against it. It is true that SLO.Growth looks very similar to MGPs when it comes to relationships with certain teacher-level covariates and in terms of intertemporal stability, but they clearly do not convey the same information about student growth. Even when focusing on student achievement results in the same subject domain, the observed correlation between SLO.Growth and MGPs is weak—the highest was $r = 0.13$ in ELA and $r = 0.29$ in Math. It is important to appreciate that SLO.Growth represents an attempt at a criterion-referenced growth measure, while an MGP is purely norm-referenced. Although one might expect to see that when students in a classroom all score higher on a test than peers across the district with similar prior year scores, that these students will have also shown growth in a criterion-referenced sense, this need not be the case. Conversely, if all students in a classroom are showing the same evidence of criterion-referenced growth on their SLOs, a teacher's MGP might still be close to the district average if it is generally the case that students across the district are all showing similar amounts of criterion-referenced growth. The results from this study suggest that when both SLO.Growth and an MGP are available that they should not be regarded as interchangeable. Whether it is defensible to use SLO.Growth as an alternative growth indicator when an MGP is not available is still an open question given the restrictions we placed on our analytic samples.

Ultimately, the validity of SLOs and the indicators that derive from them will depend to a great extent upon the validity of the assessments that underlie SLO preparedness and end-of-year command levels. What constructs do these assessments measure? What is the underlying theory of student cognition? What design principles were used to write tasks and items? When performance-tasks are used, to what extent are the scores generalizable? And so on. There is considerable evidence that can be brought to the table for PARCC (and other large-scale assessments) in this regard. The validity of the many different assessments that factor into SLO classifications is in this sense both a black box and an open question. In other words, while we would not argue that the assessments currently used in places like DPS for purposes of SLO classifications are necessarily invalid (indeed, the open-ended tasks shown in Figure 1 have some very promising features), very little systematic or formal evidence has been gathered and made publicly available in this regard. As such validity evidence is critical, the limited capacity of school district staff to collect and present such evidence represents a major impediment to the long-term dual purpose use of SLOs that has been envisioned.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. Educational Measurement: Issues and Practice, 28(4), 42–51.

Briggs, D. C. (2016). Can Campbell's Law be mitigated? In H. Braun (Ed.), *Meeting the Challenges to Measurement in an Era of Accountability*. NCME Books Series, Routledge.

Campbell, D. T. (1976). Assessing the impact of planned social change. The Public Affairs Center, Dartmouth College, Hanover New Hampshire, USA.

Castellano, K. E., & Ho, A. D. (2013). A Practitioners Guide to Growth Models. Council of Chief State School Officers.

Doherty, K. M., & Jacobs, S. (2013). State of the states 2013: Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice. National Council on Teacher Quality. http://www.nctq.org/dmsView/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and

Teacher-Level Models in Missouri. *Statistics and Public Policy*, *1*(1), 19–27. https://doi.org/10.1080/2330443X.2013.856152

Hall, E., Gagnon, D., Schneider, M.C., Marion, S., Thompson, J., (2014). State Practices Related to the Use of Student Achievement Measures in the Evaluation of Teachers in Non-Tested Subjects and Grades. Unpublished manuscript. Retrieved from: http://www.nciea.org/publication_PDFs/Gates%20NTGS_ Hall%20082614.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). How states use student learning objectives in teacher evaluation systems: a review of state websites (REL 2014–013). Washington DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Long, J. Scott. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: SAGE Publications.

McCaffrey, D. F, Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, *4*(4), 572–606.

Morganstein, D., & Wasserstein, R. (2014). ASA Statement on Value-Added Models. *Statistics and Public Policy*, *1*(1), 108–110. https://doi.org/10.1080/2330443X.2014.956906

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S, (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39, 369-393.