Data-based student learning objectives for teacher evaluation

Shuqiong Lin, Wen Luo, Fuhui Tong, Beverly J. Irby, Rafael Lara Alecio, Linda Rodriguez and Selena Chapa

Abstract: Student learning objectives (SLOs) have become an increasingly popular tool for teacher evaluations as an alternative to Value-added Models (VAMs). However, the use of SLOs faces two major challenges. First, the target setting is mostly subjective and arbitrary. Second, there is little evidence on the reliability and validity of the tool. In this paper, we proposed three data-based SLO target-setting models: split, banded, and class-wide models. The data-based approach ensures that the targets set for students are challenging yet realistic and achievable. Using data of 176 pre-kindergarten teachers and two cohorts of students from a large school district in Texas, we investigated the reliability and predictive validity of teachers’ SLO scores. Results indicated that teachers’ SLO scores had moderate to high consistency across different subtests, and moderate stability over time. Teachers’ SLO scores were also demonstrated to be useful in predicting future students’ achievement, which supported the predictive validity of the tool.

Keywords: Predictive validity; reliability; student learning objectives; target setting; teacher evaluation

ABOUT THE AUTHOR
Shuqiong Lin is currently working at American Institutes for Research as a quantitative researcher. She has earned her PhD in Research, Measurement, and Statistics from Texas A&M University. Dr Lin’s research interest lies in the area of applying quantitative methods to improve educational evaluations and assessments. She is currently working on various research projects sponsored by the National Evaluation of Investing in Innovation (i3), Institute of Education Sciences (IES), and Supporting Effective Educator Development Grant (SEED).

PUBLIC INTEREST STATEMENT
Over the past decade, U.S. educational agencies have made great efforts to develop and implement rigorous, transparent, and equitable teacher evaluation systems. Recently, student learning objectives (SLOs), a tool for setting learning targets and guiding students’ academic growth, have been used widely in the US for teacher evaluations. SLO is flexible, easy to implement, and useful for instructional planning. However, in order for SLO to be a valuable measure within teacher evaluation systems, there needs to be quality control as well as evidence on the validity and reliability of the tool. This study offers educators clear guidelines on how to set rigorous data-based targets. Data-based SLOs showed moderate levels of consistency across different subtests and stability over time. Teachers’ SLO scores were also found to be useful in predicting future students’ achievement. These findings supported the use of SLO as a medium-stakes teacher evaluation tool.
1. Introduction

Good teachers can positively affect not only students’ academic learning but also students’ future success (e.g., Makkonen, 2013; Rockoff, 2004). Relatedly, over the past decade, the U.S. Department of Education and educational organizations have made great efforts to develop and implement rigorous, transparent, and equitable teacher evaluation systems. One such system that has attracted great attention is the use of a student growth-based measurement as a significant indicator of teacher performance (Marzano & Toth, 2013). The quest of tools that can link teachers’ performance to students’ academic growth has long been pursued (e.g., Chetty, Friedman, & Rockoff, 2014; Darling-Hammond & Youngs, 2002; Kimball, White, Milanowski, & Borman, 2004; Sanders & Horn, 1994). One tool that has been adopted by many educational agencies is the value-added models (VAMs). VAMs avoid subjective ratings (Murphy, 2012) and can isolate the effects of non-educational factors such as students’ family background (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Raudenbush, 2004). However, VAMs require using high-stakes standardized tests which exclude 70% of the teachers for subjects with no standardized measurement, such as art (Chapman, 2014; Gill, Bruch, & Booker, 2013). VAMs are based on complicated statistical models that require strong model assumptions (McCaffrey et al., 2003; Reardon & Raudenbush, 2009) and are not easily understood by stakeholders such as teachers and administrators. Therefore, the use of VAMs as a high-stakes teacher evaluation tool is highly controversial (e.g., Amrein-Beardsley & Holloway, 2019).

Recently, an alternative tool, student learning objectives (SLOs), is continuously gaining popularity. SLOs measure students’ academic gains and assess teacher performance based on the degree to which pre-set targets are attained (Community Training and Assistance Center, 2013; Race to the Top Technical Assistance, 2010). Compared to VAMs, SLOs have several distinct advantages. First, SLOs can be applied to either formal test (e.g., high-stakes test) or informal assessments (e.g., locally developed tests or classroom assessments); second, SLO results can be easily explained to and understood by people with limited statistic background; third, SLOs can be used in tandem with instructional planning by allowing teachers to set up the goal and implement towards the goal. However, in order for SLOs to be a valuable measure within teacher evaluation systems, there needs to be evidence regarding the validity and reliability of the tool. So far, little research has examined the reliability and validity of using SLO scores for teacher evaluations. In addition, there are concerns regarding the target-settings in SLOs because they are mostly subjective and lack of scientific references (Marion, DePascale, Domaleski, Gong, & Diaz-Biello, 2012).

To address these issues, we had two specific purposes for our research. First, we proposed data-based target-setting guidelines (i.e., data-based SLOs) that use existing student data to inform the target-setting process. Second, we provided statistical evidence about the reliability and validity of the data-based SLOs. In the following sections, we first introduced the background of SLOs for teacher evaluation. Then, we described the proposed data-based target-setting guidelines and illustrate the application of the guidelines with a group of pre-kindergarten teachers in a large school district in Texas. With the data from the same school district, we tested two types of reliabilities: (a) the consistency of teachers’ data-based SLO scores across different subtests and (b) the year-to-year stability of teachers’ data-based SLO scores. Finally, we examined the predictive validity of teachers’ data-based SLO scores in predicting future students’ achievement.

2. Literature review

SLOs were initially used in K-12 in Denver, Colorado, as a new type of teacher evaluation measure to determine pay-for-performance (Joyce, Harrison, & Murphy, 2016). In a recent review, Lacireno-Paquet, Morgan, and Mello (2014) summarized that more than 30 states have included or have at least piloted SLOs in their teacher evaluation systems. Maryland, Michigan and Arizona required SLOs to be used for all teachers, and a few other states applied SLOs for specific subgroups, such as teachers in non-tested subjects (e.g., Mississippi and Georgia states). In a more recent review of teacher evaluation tools, Gagnon, Hall, and Marion (2017) indicated that some type of SLO usage is in place in approximately 66% of the states. They also indicated that some state-level education agencies have oversight of the SLO procedures, while others leave it to the local school districts to monitor.
2.1. Procedure of SLOs

Generally, applying SLOs for teacher evaluation involves three steps: (a) setting targets, (b) assessing student growth, and (c) evaluating teachers based on students’ target-reaching condition. In the target-setting stage, teachers or project leaders set measurable and reasonable post-test (i.e., end-of-course test) target scores for students at the beginning of the course based on students’ pre-test (i.e., beginning-of-course test) performance or any available trend data. During the instructional period, teachers reflect on students’ academic progress and classroom practice and strive to help students to achieve the target scores. At the end of the course, students whose post-test scores are equal to or higher than their target scores are counted as reaching targets. Students whose post-test scores are lower than their target scores are counted as failing to reach targets. Finally, a teacher’s SLO score is calculated as the proportion of his/her students who are defined as reaching targets over all students she/he teaches. These teachers’ SLO scores are collected and typically used for teacher categorization (Austin Independent School District, 2014; Barge, 2013).

Currently, there are three commonly used target-setting models when applying SLOs: split, banded, and class-wide models (Center for Assessment, 2014; EngageNY, 2013). Some states, such as New York, Georgia and Missouri, allow school districts to choose their own target-setting models among those three (Georgia Department of Education, 2013; Missouri Department of Elementary and Secondary Education, 2014; New York State Education Department, 2013). Other states apply one or two specific models from those three. For example, in Ohio State, teachers are asked to use class-wide target-setting models. If students show a wide range of skills and ability, banded target-setting models are used to replace class-wide target-setting models (Ohio Department of Education, 2013). Teachers at Utah are guided to apply split target-setting models (Utah State Office of Education, 2014).

Under the split model, a unique target is set for each student based on his/her pre-test score. Specifically, the target post-test score \( T_{\text{split}} \) is calculated using Equation (1) as follows:

\[
T_{\text{split}} = X + (\psi - X) \cdot \phi
\]

where \( X \) is the pre-test score, \( \psi \) is the maximum points of a test, and \( \phi \) is the split parameter, representing the proportion of the target gain score out of the maximum possible gain score. A commonly used but arbitrarily chosen value for \( \phi \) is 0.5 (also called half-the-distance growth targets). This means the target gain score for a student is half of the maximum possible gain score for the student.

The banded model is used when there are tiered levels of expectations for students. Specifically, students are grouped into different bands according to their pre-test scores, and within each band a common target is set. For example, students who have pre-test scores lower than 50 are required to reach 60 in their post-test; students who have pre-test ranging from 51 to 70 are required to reach 80, and so on. The banded model generally uses an arbitrarily determined target for each band.

The class-wide model sets the same target for all students; therefore, it is not individual specific and does not require information about student pre-test scores. The shared target can be set according to the need of a course. For instance, the target can be either high to represent the mastery level of achievement or low to represent the passing level of achievement (New York State Education Department, 2013).

2.2. Advantages of SLOs

SLOs have several advantages. First, SLOs can be applied to both tested and non-tested subjects using either formal (nationwide or state-based test) or informal (locally developed tests or classroom assessments) academic performance scores (Bergin, 2015); thus, allowing the SLO tool to be used in any grade and any subject. Second, SLOs allow various target-setting methods to accommodate the diverse situations and teachers’ needs across subjects and grades. Third, SLOs are implemented directly by teachers and campus principals who can be actively involved in setting
learning targets and evaluating learning growth based on simple rubrics. Thus, SLOs are easier to understand for people with limited statistic background than other teacher evaluation tools which often require complicated statistic modeling (Bergin, 2015). Fourth, SLOs can be used in tandem with instructional planning and improvement (Gill et al., 2013). It has been demonstrated that SLOs can positively affect students’ achievement by providing their teachers with clear structures of what knowledge and skills the students need to gain in courses (Aziz, Yusof, & Yatim, 2012). The use of SLOs may also increase students’ effort and self-motivated participation (Paolini, 2015) and has been shown to improve their learning outcomes in math, reading, and writing (Briggs, Diaz-Bilello, Moul, Turner, & Bibilos, 2014). Due to these advantages, teachers, especially new teachers, have claimed that using SLOs based teacher evaluation encourages them to analyze student growth data and therefore improve their teaching (Lamb, Schmitt, Gross, & Cornetto, 2013; Schmitt, 2014).

2.3. Critiques of SLOs
Despite these merits, the wide use of SLOs in teacher evaluation systems has been limited for two main reasons. First, little is known about the reliability and validity of using SLO scores for teacher evaluation (Bergin, 2015; Gill et al., 2013). Given the fact that SLOs have been widely applied, it is necessary and pressing to shift from claims-based studies to validation-based studies in which researchers provide scientific evidence about how reliable and valid teachers’ SLO scores are. Crouse, Gitomer, and Joyce (2016) claimed that there are strong concerns about the validity, reliability, and accuracy of SLO scores, especially given variability in current levels of score calibration, assessment design, and quality controls. However, they noted that SLOs can be a valuable measure within teacher evaluation systems if quality controls are in place.

According to our review, to date, no researchers have investigated the reliability of SLOs for teacher evaluation, and only limited reported studies have included validation-based explorations for SLOs. For example, researchers in the Community Training and Assistance Center (2004, 2013) found that teachers’ SLOs ratings were positively associated with individual students’ academic achievement in elementary and middle school. In a series of studies, Schmitt (2011, 2014), Schmitt, Cornetto, Lamb, and Imes (2009), and Schmitt and Ibanez (2011) as well as Proctor, Walters, Reichardt, Goldhaber, and Walch (2011) reported similar results but at the aggregated school level as opposed to the individual student level. Only a few researchers (e.g., Balch & Springer, 2015; Community Training and Assistance Center, 2013; Goldhaber & Walch, 2011) have provided direct evidence about the convergent validity of SLOs and found significant positive associations between SLOs and other teacher evaluation tools. With such limited examinations of reliability and validity of SLOs, precision, fairness, and stability of the estimated teaching effectiveness based on SLOs are questionable.

Second, since target-settings in SLOs are subjective, teachers are lack of scientific references for setting appropriate target scores for students (Marion et al., 2012). SLO target scores are driven by the needs of the students assigned to teachers; thus, it is reasonable to ask teachers who interact with students directly to set goals. Among the 24 states that have adopted the target-setting models described above, 18 states involve teachers in setting their own SLOs (Lacireno-Paquet et al., 2014). Some teachers have complained about others gaming the system by setting low targets to get better evaluation scores (Bergin, 2015; Gill et al., 2013). This gaming issue is noted by some researchers (e.g., Balch & Springer, 2015) for explaining the low validity of using SLOs found in their studies. To address this problem, many districts have required targets to be approved by other teachers, who teach the same subjects, principals, or district-level experts (Gill, English, Furgeson, & McCullough, 2014). However, completely relying on the subjective judgments of teachers and principals still might lead to biased target-settings and undermine the fairness of teacher evaluation. Therefore, clear guidelines for setting achievable and reasonable target scores are needed when using SLOs in the assessment of teacher performance.
In summary, SLOs have several distinct advantages: flexible, straightforward to implement, easy to understand, and useful for instructional planning. However, the subjective and arbitrary way of target-setting in practice and the limited evidence for reliability and validity may cause potential risks and limit the wider usage of SLOs for teacher evaluations. Thus, in this study, we proposed data-based target-setting guidelines (i.e., data-based SLOs) that provide scientific reference for setting appropriate target scores and examined the reliability and validity of the proposed data-based SLOs.

3. Method

3.1. Participants
Ethical approval for this study was obtained from the participating university and schools. A total of 176 pre-kindergarten (pre-K) teachers from eight pre-K schools in a large urban school district located in Texas participated in the study. The teachers taught bilingual education, English as a second language (ESL), or regular education programs. During the 2-year study period, the teachers taught two cohorts of students. The first cohort (i.e., 2014–2015 academic year) consisted of 3,311 pre-K students who completed tests. The average age of the students was 57.75 months ($SD = 3.58$). Among these students, 46.64% enrolled in bilingual programs, 7.94% in ESL programs, and 45.21% in regular education programs. The majority of the students were Hispanic (72.18%), followed by African Americans (24.58%). Female students represented 51.56% of this sample as noted in Table 1.

Among those 176 teachers who taught the cohort-1 students, 164 of them participated in the study the following year (i.e., 2015–2016 academic year). The demographics of the second cohort of students ($N = 3,213$) closely approximated those of the first cohort of students as noted in Table 2.

3.2. Measures
Data-based SLOs can be applied using either formal (i.e., nationwide or state-wide test) or informal (i.e., locally developed tests) academic performance test scores. In this study, we used the Bracken

| Table 1. Demographics and BBCS:E test scores of students ($N = 3,311$) in cohort-1 |
|-----------------|---------|---------|---------|---------|---------|---------|
| Gender          | N       | %       | Pre-test | Pre-test SD | Post-test | Post-test SD |
| Female          | 1707    | 51.56   | 40.34    | 11.85     | 52.59    | 13.35     |
| Male            | 1604    | 48.44   | 38.08    | 11.65     | 50.52    | 13.70     |
| Ethnicity       |         |         |         |         |         |         |
| Asian           | 32      | 0.97    | 38.06    | 12.58     | 49.47    | 13.09     |
| Black           | 814     | 21.58   | 43.67    | 12.02     | 51.89    | 12.88     |
| Hispanic        | 2390    | 72.18   | 37.62    | 11.28     | 51.46    | 13.76     |
| American Indian | 4       | 0.12    | 38.00    | 12.30     | 47.75    | 4.50      |
| Multirace       | 21      | 0.63    | 44.90    | 12.08     | 52.52    | 13.74     |
| Pacific Islander| 3       | 0.09    | 43.00    | 1.73      | 58.33    | 7.51      |
| White           | 47      | 1.42    | 42.79    | 12.31     | 53.40    | 15.65     |
| Program         |         |         |         |         |         |         |
| Bilingual       | 1551    | 46.84   | 36.89    | 10.30     | 52.70    | 13.63     |
| ESL             | 263     | 7.94    | 32.06    | 11.38     | 44.12    | 13.18     |
| Regular         | 1497    | 45.21   | 42.80    | 12.02     | 51.72    | 13.19     |
| Overall         | 3900    | 11.68   | 51.58    | 13.50     |           |           |
Basic Concept Scale: Expressive (BBCS:E; Bracken, 2006), which is a norm-referenced standardized assessment used to evaluate the acquisition of basic concepts nonverbally for 3- to 6-year-old students. These basic concepts are categorized and measured by six subtests: School Readiness Composite (SRC), Direction/Position, Self-/Social Awareness, Texture/Material, Quantity, and Time/Sequence. SRC subtest measures a child’s overall knowledge of colors, letters, numbers/counting, sizes/comparisons, and shapes. Direction/Position measures a child’s acquisition of order and position of objects. Self-/Social Awareness defines a child’s ability of understanding relationships and roles and taking the perspectives of others and applying it to the interactions with other people. Texture/Material measures a child’s ability of recognizing the physical characteristics of objects. Quantity measures a child’s ability of measuring object by height, weight, and length. Time/Sequence tests a child’s ability of understanding the order and sequence of objects. These subtest areas are strongly related to the cognitive and language development for early school achievement. Following the manual, we converted raw scores of each subtest into scaled scores, which were summed as scaled composite scores ranging from 6 to 114. Based on the scaled scores, the BBCS:E provides four classifications (i.e., very delayed, delayed, average, advanced). Good test–retest reliabilities (0.80 < r < 0.95) and high inter-rater agreement (0.96 < ρ < 0.99) were reported for the BBSC:E subtests (Bracken, 2006).

In our study, the BBCS:E was administrated across the two cohorts. For the first cohort, BBCS:E was tested in September 2014 as pre-test and May 2015 as post-test. For both pre- and post-test, every student has six scaled subtest scores, one scaled composite score, and corresponding classifications. These scaled scores and classifications were utilized for all further analyses. Table 1 lists students’ means (M) and standard deviations (SD) of pre and post scaled composite scores. At the beginning of that academic year, students’ mean test scores equaled 39.00 (SD = 11.68) with the corresponding descriptive classification as Delayed. The mean post-test scores equaled to 51.58 (SD = 13.50) with the classification as Average. For the second cohort, the same BBCS:E was administered in September 2015 as pre-test and May 2016 as post-test. Students’ mean pre- and post-test BBCS:E scaled composite scores are shown in Table 2, which are similar to those students in cohort-1.

| Table 2. Demographics of students (N = 3,213) in cohort-2 |
|-------------|------------|----------------|----------------|
| Gender      |            | Pre-test M     | Pre-test SD    |
| Female      | 1598       | 38.67          | 11.67          |
| Male        | 1615       | 36.78          | 11.83          |
| Ethnicity   |            | Post-test Mean | Post-test SD   |
| Asian       | 32         | 34.79          | 13.17          |
| Black       | 814        | 43.09          | 11.55          |
| Hispanic    | 2390       | 35.76          | 11.13          |
| American Indian | 4 | 24.50         | 12.03          |
| Multirace   | 21         | 40.93          | 12.13          |
| Pacific Islander | 3 | 44.00         | 2.83           |
| White       | 47         | 45.97          | 13.85          |
| Program     |            | Overall        |                |
| Bilingual   | 1472       | 36.89          | 10.30          |
| ESL         | 347        | 32.06          | 11.38          |
| Regular     | 1394       | 42.80          | 12.02          |
| Overall     | 39.00      | 11.68          | 51.58          | 14.25          |

Lin et al., Cogent Education (2020), 7: 1713427 https://doi.org/10.1080/2331186X.2020.1713427
4. Data-based target-setting guidelines

In this section, we utilize cohort-1 students’ scaled composite scores from BBCS:E to illustrate the data-based target-setting guidelines. Given the nature of the different programs, i.e., bilingual education (in which native Spanish-speaking students were taught in both Spanish and English), ESL education (in which native Spanish-speaking students were provided structured support in developing English proficiency), and regular education, we apply the data-based SLOs for each program separately.

4.1. The split model

When applying the general split models, $T_{\text{split}} = X + (\psi - X) + \varphi$, a commonly used but arbitrarily chosen value for $\varphi$ is 0.5. For instance, if a bilingual student’s pre-test score was 45 and the maximum point of the test was 114; then, his/her target score was set to be 79.50 [i.e., 45 + (114–45)*0.50 = 79.50]. With such arbitrarily chosen proportions, the target gains could be set too high or too low, causing teachers’ SLO scores to be highly skewed and less discriminating between effective and ineffective teachers. To avoid such an issue, we propose the data-based target-setting guideline that sets the value of $\varphi$ based on past student data. To compute the split parameter, we first obtained the mean pre-test ($\mu_X$) and the mean post-test scores ($\mu_Y$) based on all previous students who were in the same school district and took the same test. Then, $\varphi$ is computed as the ratio of the actual mean growth over the maximum possible mean growth in Equation (2):

$$\varphi = \frac{\mu_Y - \mu_X}{\psi - \mu_X}$$

(2)

Ideally, $\mu_X$ and $\mu_Y$ are obtained based on students from previous cohorts. However, if such previous test scores are not available, the current cohort of students’ pre- and post-test scores can be used instead.

The advantage of using this data-based split model is that the target is individual specific. Individual students’ prior achievement is taken into consideration when setting the goals and different individuals have different target scores. However, it requires that both the pre-test (i.e., beginning-of-course test) and post-test (i.e., end-of-course test) achievements are available and should be measured on the same scale. When only the current student data are available, the drawback of using current students’ data is that the target scores can only be set and known at the end of courses when post-test scores are available, which means teachers can still be evaluated, but SLOs would not be able to be used for instructional planning.

In this study, because previous data were not available for target setting for cohort-1 students, we used their own pre- and post-test scores for target setting. According to the pre- and post-test means for each of the three education programs and the maximum scaled composite score (i.e., 114), we computed the split parameter as $\varphi = 0.20$ for bilingual program, $\varphi = 0.15$ for ESL program, and $\varphi = 0.13$ for regular program based on Equation (2). Then, each student’s target score is set using Equation (1). For example, if a bilingual student’s pre-test score was 45, then his/her target score was set to be 58.80 [i.e., 45+(114–45)*0.20 = 58.80].

4.2. The banded model

When applying the banded model, students are first grouped into different bands according to their pre-test scores, and within each band a common target is set. In the data-based target-setting approach, we proposed to use the mean post-test score $\mu_Y$ from previous students in each band as the target. If no previous data are available, the mean of post-test scores based on current students can be used, which has the same drawback as previously mentioned. Like the split model, banded models require information of student pre-test scores. However, it does not require the pre- and post-test scores to be on the same scale, which makes it more flexible than the split model. The banded model can be easily applied to any tests that classify and report students’ knowledge on an ordinal scale. In addition, the banded model is more suitable for teachers whose students show a wide range of skills and abilities in the target subject.
As an illustration, we used the normed classifications provided by the BBCS:E test to group students into four bands: Very Delayed (pre-test scores <32), Delayed (pre-test scores between 32 and 46), Average (pre-test scores between 47 and 73), and Advanced (pre-test scores >74). Since no previous data existed, we computed the mean post-test scores within each band for the same cohort of students, which were used as the target scores. We applied this target-setting within each education program. For the bilingual education program, we used the following targets:

\[
\text{Band 1 (Very Delayed): } X_i < 32 \quad \text{Target} = 42.94 \\
\text{Band 2 (Delayed): } 32 \leq X_i \leq 46 \quad \text{Target} = 54.73 \\
\text{Band 3 (Average): } 47 \leq X_i \leq 73 \quad \text{Target} = 64.99. \\
\]

For the ESL education program, the following targets were used:

\[
\text{Band 1 (Very Delayed): } X_i < 32 \quad \text{Target} = 36.34 \\
\text{Band 2 (Delayed): } 32 \leq X_i \leq 46 \quad \text{Target} = 49.53 \\
\text{Band 3 (Average): } 47 \leq X_i \leq 73 \quad \text{Target} = 62.04. \\
\]

It should be noted that no bilingual or ESL students were classified as Advanced based on their pre-test achievement. For regular education program students, the following targets were used:

\[
\text{Band 1 (Very Delayed): } X_i < 32 \quad \text{Target} = 36.84 \\
\text{Band 2 (Delayed): } 32 \leq X_i \leq 46 \quad \text{Target} = 49.27 \\
\text{Band 3 (Average): } 47 \leq X_i \leq 73 \quad \text{Target} = 61.19 \\
\text{Band 4 (Advanced): } X_i > 74 \quad \text{Target} = 78.13. \\
\]

4.3. The class-wide model

The class-wide model sets the same target for all students. Generally, the target scores are set to represent the mastery or passing level of achievement. However, sometimes it is hard to clearly define the mastery or passing level of achievement and choose an appropriate score to represent the corresponding level. In this case, we propose the data-based target-setting guideline which uses the mean of post-test scores from previous students as the common target. In this study, the class-wide target was set as 52.70, 44.12, and 51.72 for bilingual, ESL and regular program students, respectively.

It should be noted that, for the banded and class-wide models, there exist cases in which students’ pre-test scores already exceed their target scores. For those students, as long as they can maintain their achievement and have post-test scores higher than their corresponding target scores, they are still qualified as target-reaching students.

5. Application of the data-based target-setting guidelines

The data-based target-setting guidelines provide teachers and principals with a useful reference because targets that are set following the data-based guidelines are achievable and reasonable in general. For the majority of students, these targets are appropriate. However, there may be situations in which the data-based targets cannot be used directly, such as exceptions for student absenteeism, or student mobility. In these cases, teachers have the flexibility to modify the SLO targets. States and districts have crafted business rules to guide the process of target modification in case of student absenteeism, mobility, extended leave of teachers, assignment change, and students with disabilities.
(Potemski, 2013). However, the implementation of these rules for target modifications is out of the scope of this study.

In this study, to control for the influence of target modifications, the data-based targets were used directly without modifications. Students’ post-test scores were compared with their data-based target scores and students were labeled as either target-reaching or non-target-reaching. Teachers’ SLO scores were computed as the proportion of target-reaching students. Three sets of teachers’ SLO scores were estimated according to the three different data-based target-setting models. Results showed that teachers’ data-based SLO scores had similar means and standard deviations across the three data-based target-setting models, i.e., $M = 0.47, SD = 0.28$ for split, $M = 0.48, SD = 0.26$ for banded, and $M = 0.47, SD = 0.26$ for class-wide (see Figure 1 for the distributions of teachers’ SLO scores).

Figure 1. Histograms of teachers’ SLO scores by different data-based target-setting models.
6. Reliability of data-based SLOs for teacher evaluation

We examined the reliability of teachers' SLO scores from two aspects: consistency across different subtests and stability over the years.

6.1. Consistency across different subtests

As mentioned before, the BBCS:E test consists of six subtests assessing students' achievement in six areas. Because teachers' instruction affects all six areas, it is hypothesized that teachers' SLO scores estimated based on the six subtests should be moderately correlated. This is analogous to internal consistency reliability that shows consistency of results between different items or subtests within a particular test (Trochim & Donnelly, 2001).

6.1.1. Participants and measures

We used the six BBCS:E subtest scaled scores of cohort-1 for this analysis. Table 3 lists the pre-and post-test mean scaled scores of each subtest by education program type. The overall mean is the mean scaled scores of all students in each program, while the banded mean is the mean scores of students within each normed band.

6.1.2. Analysis

Using the data-based target-setting models, we set targets for each subtest. Table 4 shows the data-based split parameters and target scores for each subtest by programs. Teachers' SLO scores were then computed for each subtest separately and the Pearson $r$ correlations were computed among the six sets of teachers' SLO scores.

6.1.3. Result

Table 5 lists the means of teachers' SLO scores across the six subtests by data-based target-setting models. Table 6 contains all the correlations among subtests under each data-based target-setting model. As shown in Table 6, in general, there are moderate to high correlations among teachers' SLO scores among different subtests (ranging between .41 and .74), regardless of target-setting models.

6.2. Stability across years

Because teachers' effects are unlikely to change dramatically from year to year, we hypothesize that teachers' SLO scores will be moderately correlated between cohort-1 and cohort-2, which will provide evidence to the test–retest reliability of teachers' SLO scores.

6.2.1. Participants and measures

The 164 pre-K teachers who were evaluated in both academic years and the two cohorts of students were included. The students' scaled composite scores from BBCE:E were utilized for this analysis.

6.2.2. Analysis

The data-based split parameter, banded target scores and class-wide target scores derived from the cohort-1 data were used as targets for computing teachers' SLO scores in both cohort-1 and cohort-2. Pearson $r$ correlations of teachers’ SLO scores between cohort-1 and cohort-2 were computed.

6.2.3. Results

Table 7 includes the mean SLO scores for teachers which increased slightly in cohort-2 (Mean = .53) compared to cohort-1 (Mean = .47). The correlations of teachers' SLO scores between the 2 years were $r_{split} = .48$, $r_{banded} = .44$ and $r_{class-wide} = .40$, indicating moderate stability over time.

7. The predictive validity of SLOs

Increasing student academic achievement is the primary goal of schooling and one of the most important criteria for judging teachers' effectiveness. Researchers, therefore, have examined the correlation between teachers' evaluation scores and students' achievement scores for validating teacher evaluation results (e.g., Hill, Kapitula, & Umland, 2011; Schmitt & Lamb, 2014). If teachers'
### Table 3. Mean subtest scores by program type based on cohort-1 (N = 3,311)

|                  | School readiness composite | Direction/position | Self-/social awareness | Texture/material | Quantity | Time/sequence |
|------------------|---------------------------|--------------------|------------------------|-----------------|----------|---------------|
|                  | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test |
| **Bilingual**    |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| Overall          | 5.35     | 11.61    | 6.1      | 9.17     | 6.67     | 8.78     | 6.35     | 8.91     | 7.13     | 10.89    | 5.27     | 7.19     |
| Banded 1         | 3.13     | 9.88     | 3.31     | 7.14     | 3.01     | 6.84     | 3.66     | 6.67     | 3.09     | 8.06     | 2.72     | 3.37     |
| Banded 2         | 5.82     | 12.15    | 6.03     | 9.46     | 6.18     | 8.45     | 6.15     | 8.73     | 6.04     | 10.31    | 5.82     | 7.87     |
| Banded 3         | 9.13     | 13.97    | 9.21     | 11.08    | 8.77     | 9.96     | 9.09     | 11.22    | 9.21     | 12.35    | 9.01     | 13.47    |
| Banded 4         | 13.67    | 15.67    | 12.75    | 13.13    | NA       | NA       | 13.42    | 13.86    | 13.42    | 15.36    | 13.50    | 18.10    |
| **ESL**          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| Overall          | 6.02     | 11.18    | 4.58     | 6.94     | 4.77     | 7.51     | 5.75     | 7.17     | 5.85     | 3.06     | 5.07     | 6.93     |
| Banded 1         | 2.81     | 9.43     | 2.94     | 5.51     | 2.58     | 5.64     | 3.52     | 5.73     | 3.91     | 6.31     | 2.79     | 3.20     |
| Banded 2         | 5.74     | 11.10    | 5.99     | 8.47     | 5.93     | 8.58     | 6.12     | 7.34     | 5.80     | 8.18     | 6.00     | 8.18     |
| Banded 3         | 9.58     | 13.01    | 8.88     | 9.84     | 9.03     | 11.01    | 9.54     | 10.13    | 9.26     | 10.85    | 8.94     | 14.01    |
| Banded 4         | 14.20    | 16.21    | NA       | NA       | NA       | NA       | NA       | NA       | NA       | NA       | NA       | NA       |
| **Regular**      |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
| Overall          | 7.43     | 11.27    | 6.74     | 8.67     | 6.87     | 9.34     | 7.13     | 8.65     | 7.13     | 9.49     | 6.37     | 9.37     |
| Banded 1         | 3.05     | 7.78     | 3.41     | 5.99     | 2.91     | 6.50     | 3.60     | 6.20     | 3.90     | 7.21     | 2.21     | 2.36     |
| Banded 2         | 5.98     | 10.5     | 6.15     | 8.28     | 6.15     | 8.91     | 6.24     | 7.81     | 6.05     | 8.84     | 6.23     | 9.13     |
| Banded 3         | 9.64     | 12.91    | 8.94     | 10.33    | 9.11     | 10.87    | 9.39     | 10.58    | 9.11     | 10.79    | 9.24     | 14.21    |
| Banded 4         | 13.55    | 14.64    | NA       | NA       | NA       | NA       | NA       | NA       | 13.4     | 14       | NA       | NA       |
|                  | School readiness composite | Direction/position | Self-/social awareness | Texture/material | Quantity | Time/sequence |
|------------------|----------------------------|--------------------|------------------------|------------------|----------|--------------|
|                  | 0.46                       | 0.24               | 0.17                   | 0.20             | 0.32     | 0.12         |
| Bilingual Split  | 9.88                       | 7.14               | 6.84                   | 6.67             | 8.06     | 3.37         |
|                  | 12.15                      | 9.46               | 8.45                   | 8.73             | 10.31    | 7.87         |
|                  | 13.97                      | 11.08              | 9.96                   | 11.22            | 12.35    | 13.47        |
|                  | 15.67                      | 13.13              | NA                     | 13.86            | 15.36    | 18.10        |
| Class-wide       | 11.61                      | 9.17               | 8.78                   | 8.91             | 10.89    | 7.19         |
|                  | 0.40                       | 0.16               | 0.19                   | 0.11             | 0.17     | 0.12         |
| ESL Split 1      | 9.43                       | 5.51               | 5.64                   | 5.73             | 6.31     | 3.20         |
|                  | 11.10                      | 8.47               | 8.58                   | 7.34             | 8.18     | 8.18         |
|                  | 13.01                      | 9.84               | 11.01                  | 10.13            | 10.85    | 14.01        |
|                  | 16.21                      | NA                 | NA                     | NA               | NA       | NA           |
| Class-wide       | 11.18                      | 6.94               | 7.51                   | 7.17             | 3.06     | 6.93         |
|                  | 0.33                       | 0.33               | 0.20                   | 0.13             | 0.20     | 0.20         |
| Regular Split 1  | 7.78                       | 5.99               | 6.50                   | 6.20             | 7.21     | 2.36         |
|                  | 10.50                      | 8.28               | 8.91                   | 7.81             | 8.84     | 9.13         |
|                  | 12.91                      | 10.33              | 10.87                  | 10.58            | 10.79    | 14.21        |
|                  | 14.64                      | NA                 | NA                     | NA               | 14.00    | NA           |
| Class-wide       | 11.27                      | 8.67               | 9.34                   | 8.65             | 9.49     | 9.37         |
SLO scores are valid measures of teachers’ true effectiveness, we would expect that higher SLO scores are indicative of more effective teaching and that teachers with higher SLOs make more contributions to their students’ learning, thus producing larger learning growth for future students. Accordingly, we hypothesize that using teachers’ SLO scores based on cohort-1 can significantly predict cohort-2 students’ learning growth.
7.1. Analysis and results
Due to the multilevel data structure (i.e., students nested within teachers), we used the following two-level hierarchical linear model in Equation (3) as follows:

\[ \text{Level 1} : Y_{ij} = \beta_0 + \beta x_{ij} + e_{ij} \]

\[ \text{Level 2} : \beta_0 = \gamma_{00} + \gamma_01 \text{SLO}_j + u_{0j} \]

\[ \beta = \gamma_{10} \quad (3) \]

where \( Y_{ij} \) represents cohort-2 students’ end-of-year BBCS:E scores, \( x_{ij} \) represents cohort-2 students’ BBCS:E scores at the beginning of the year, and \( \text{SLO}_j \) represents teachers’ SLO scores obtained based on cohort-1 students’ data. We fit one model for each type of SLO target-setting method. Because the scales of BBCS:E scores and teachers’ SLO scores were largely different from each other, \( Y_{ij}, x_{ij} \) and \( \text{SLO}_j \) variables were all standardized before fitting this model.

The results in Table 8 show that all the \( \gamma_{01} \) values (i.e., 0.24, 0.24, and 0.22), which indicate the effect of teachers’ SLO scores on future students’ achievement, were significant (\( p \)'s < .0001). The proportional reduction in prediction error (Snijders & Bosker, 1994) when using teachers’ SLO scores to predict student achievement was .25 for the data-based split model, .24 for the data-based banded model, and .20 for the data-based class-wide model.

8. Discussion
A good teacher evaluation tool serves two main purposes: one is to help in improving teaching and the other is to assist in decision-making, including hiring, retention, and promotion. Both purposes are built on a fundamental requirement that teacher evaluation should be fairly applied to all teachers in a reliable and valid way. Growth-based teacher evaluation tools, such as value-added models, measure the impact that schools and teachers have on students’ academic progress from year to year across subjects. The measures often require complicated statistical models and only can be used with high-stakes tests; therefore, they are limited to a few subjects and teachers. As an alternative, the SLO is a simple, understandable, and user-friendly tool that involves teachers in the evaluation process and supports teachers’ work of improving student learning. Our study improves the current SLO target-setting practices by providing a set of data-based target-setting guidelines and presents reliability and validity evidence for using data-based SLOs in teacher evaluation.

8.1. Data-based vs. arbitrary target-setting approaches
The data-driven target-setting guidelines utilize past students’ test performance information to set learning objectives. Compared to arbitrarily chosen targets, data-based targets have a distinct advantage. When targets are chosen arbitrarily, they may be too easy or too hard to reach, which may result in the failure to differentiate teachers’ performance. However, when using the data-based target scores as the target-setting reference, targets are anchored at the mean achievement of the population, which ensures that the targets are achievable and realistic for the majority of students. Setting challenging yet achievable targets is also essential to keep teachers motivated, which is important for teachers’ job satisfaction and sense of fulfillment (e.g., Neves de Jesus & Lens, 2005).
8.2. Reliability and validity of teachers’ SLO scores

8.2.1. Consistency across different subtests
Teachers’ SLO scores based on different subtests of the BBCS:E had moderate to high correlations (.41 to .74), indicating moderate consistency across the BBCS:E subtests. This is expected because the six subtests cover a wide range of domains and individual teachers might have their own strong and weak areas. The moderate consistency may also be due to the variation in the alignment between curriculum and tests or the difference in test reliabilities per se (Baker et al., 2010).

8.2.2. Year-to-year stability
We demonstrated moderate-sized correlations between teachers’ SLO scores in Year 1 and Year 2, meaning that the same teachers’ SLO score could change year by year depending on the classes they taught and their accumulated teaching experience. Specifically, the data-based split model reported the highest stability with the correlation of .48, followed by the data-based banded model with a correlation of .44. The data-based class-wide model had the lowest stability with the correlation of .40. These findings are consistent with Gagnon et al. (2017) who have claimed that the SLOs based teacher evaluation differs in terms of (a) assessments selected to measure students’ growth, (b) the proportion of SLOs in teacher evaluation, and (c) the amount of guidance on the implementation of SLOs in teaching and evaluating. Moderate reliability is not only an issue for SLOs but also a general

| Table 8. Estimation of parameters by data-based target-setting model (N = 164) for predicting cohort-2 students’ achievement |
|-----------------------------------------------|
|                | Standardized coefficient | T test | p-Value |
| Split          |                               |        |         |
| Fixed effects  |                               |        |         |
| $\gamma_{00}$ | 0.01                          | 0.17   | .868    |
| $\gamma_{10}$ | 0.65                          | 53.53  | <.0001  |
| $\gamma_{01}$ | 0.24                          | 6.88   | <.0001  |
| Random effects |                               |        |         |
| $\sigma^2_u$   | 0.18                          | 8.11   | <.0001  |
| $\sigma^2_e$   | 0.39                          | 38.93  | <.0001  |
| Banded         |                               |        |         |
| Fixed effects  |                               |        |         |
| $\gamma_{00}$ | 0.01                          | 0.14   | .887    |
| $\gamma_{10}$ | 0.65                          | 53.55  | <.0001  |
| $\gamma_{01}$ | 0.24                          | 6.71   | <.0001  |
| Random effects |                               |        |         |
| $\sigma^2_u$   | 0.18                          | 8.12   | <.0001  |
| $\sigma^2_e$   | 0.39                          | 38.93  | <.0001  |
| Class-wide     |                               |        |         |
| Fixed effects  |                               |        |         |
| $\gamma_{00}$ | 0.01                          | 0.25   | .807    |
| $\gamma_{10}$ | 0.65                          | 53.45  | <.0001  |
| $\gamma_{01}$ | 0.22                          | 5.96   | <.0001  |
| Random effects |                               |        |         |
| $\sigma^2_u$   | 0.19                          | 8.15   | <.0001  |
| $\sigma^2_e$   | 0.39                          | 38.93  | <.0001  |
issue for other student growth-based teacher evaluation tools (e.g., Lockwood et al., 2007; Papay, 2011). The general lack of stability might be due to the fact that all of these evaluation tools are based on test scores that cannot capture all aspects of educational outcomes. Even if increasing students’ academic performance is the most important goal for teaching, SLOs, measures based on students’ academic growth, might oversimplify the complex nature of teaching (Cantrell & Kane, 2013; Goldhaber, Goldschmidt, & Tseng, 2013; Schmidt & Kaplan, 1971).

8.2.3. Predictive validity of teachers’ SLO scores
Teachers’ SLO scores demonstrated promising predictive validity in terms of predicting future students’ academic achievement. About 20–25% of the variance in cohort-2 students’ achievement was uniquely accounted for by teachers’ SLO scores estimated in cohort-1. These moderate contributions of teachers’ SLO scores in predicting future students’ achievement were reasonable given the many other factors that are related to students’ achievement but are not captured by teachers’ SLO scores.

8.3. Recommendations when applying data-based SLOs
For school districts that are currently using or plan to use SLOs in their teacher evaluation systems, the data-based target-setting guidelines can be easily adopted for tested subjects. For non-tested subjects, such as music and arts, districts need to first develop common performance-based assessments. When utilizing the data-based target-setting guidelines, teachers and principals should be aware that these data-based targets are appropriate for the majority of students, but there may be exceptions that warrant additional considerations, such as student absenteeism or mobility. In those cases, teachers and principals can use the data-based targets as a reference and make suitable modifications.

No one target-setting model outperforms other models across all situations. Each target-setting model has its own merits and is suitable for different assessments and needs of courses. In terms of reliability, the three target-setting models had similar consistency across different subtests. However, the data-based split model is recommended when evaluating teachers using the same test across multiple cohorts, since the data-based split model is found to have the highest stability across two years in our study. Furthermore, the data-based split model and the data-based banded model are more useful for predicting students’ learning growth in the future.

Based on the evidence for reliability and predictive validity, we believe that teachers’ SLO scores do reflect their performance to some extent and could provide useful information for personnel decisions. However, school leaders should be aware that teachers’ SLO scores could be affected by tests and vary from year to year with moderate evaluation consistency and stability found in this study. Therefore, when applying SLOs for teacher evaluation, first, we should caution against making high-stakes personnel decisions solely based on teachers’ single-year SLO scores on a single test. Researchers claimed that the higher the stakes attached to the SLOs, the more stressful the SLOs were for the teachers (Longchamp, 2017). Instead, low- or medium-stakes use of SLOs might be a better option. For example, leaders can gather information from multiple sources including classroom observations and student/parent ratings of teachers across multiple years and combine this information with SLO results to gain a more comprehensive and accurate evaluation of teaching effectiveness. Second, we should create and improve the SLO training for leaders and teachers. When participants received little training, they had relatively low engagement (Longchamp, 2017) and tended to produce poor-quality SLOs and implemented SLOs carelessly (Joyce et al., 2016), which may decrease the reliability of validity of using SLOs for teacher evaluation. Therefore, extensive training including SLOs-related theories, SLO procedures (e.g., scientific target settings), and effective teaching strategies based on the feedback obtained from applying SLOs is highly recommended.

9. Limitations and future study
The findings should be interpreted in light of the limitations of the study. First, we focused our study on pre-K teachers. It is not clear whether the results are generalizable to teachers of other
grades. Second, we only had 2 years of data; therefore, we are not able to assess the long-term stability of data-based SLOs beyond the 2 years. Third, we did not examine other types of validity, such as convergent and divergent validity. More empirical research is needed in this area to better inform the use of data-based SLOs as tools for teacher evaluation.

Funding
The authors received no direct funding for this research.

Author details
Shuqiong Lin1
E-mail: slin@air.org
Wen Luo2
E-mail: vluo@tamu.edu
Fuhui Tong1,2
E-mail: fuhuitong@tamu.edu
ORCID ID: http://orcid.org/0000-0003-0555-892X
Beverly J. Irby1,2
E-mail: irbyb@tamu.edu
Rafael Lara Alecio2
E-mail: a-lara@tamu.edu
ORCID ID: http://orcid.org/0000-0001-5007-3580
Linda Rodriguez3
E-mail: rodguezr@aldineisd.org
Selena Chapa1
E-mail: scarita@aldineisd.org
1 American Institutes for Research, USA.
2 Department of Educational Psychology, Texas A&M University, College Station, TX, USA.
3 Center for Research and Development in Dual Language and Literacy Acquisition (CRULLA), Texas A&M University, College Station, TX, USA.
4 Assistant Superintendent of Human Resources, Aldine Independent School District, Houston, TX, USA.

Citation information
Cite this article as: Data-based student learning objectives for teacher evaluation, Shuqiong Lin, Wen Luo, Fuhui Tong, Beverly J. Irby, Rafael Lara Alecio, Linda Rodriguez & Selena Chapa, Cogent Education (2020), 7: 1713427.

References
Amrein-Beardsley, A., & Holloway, J. (2019). Value-added models for teacher evaluation and accountability: Commonsense assumptions. Educational Policy, 33 (3), 516–542. doi:10.1177/0895904817719519
Austin Independent School District. (2014). Reach SLO Manual 2013-1014. Retrieved from http://www.austi
nisd.org/sites/default/files/dept/reach/SLO_Manual_2013-2014FinalRevisedJ_0.pdf
Aziz, A. A., Yusof, K. M., & Yatim, J. M. (2012). Evaluation on the Effectiveness of Learning Outcomes from Students’ Perspectives. Procedia-Social and Behavioral Sciences, 56, 22–30. doi:10.1016/j.pssb.2012.09.628
Baker, E. L., Barton, P. E., Darling-Hammond, L., Hoertel, E., Ladd, H. F., Linn, R. L., … Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper #278. Economic Policy Institute. Retrieve from http://www.epi.org/publication/bp278
Balch, R., & Springer, M. G. (2015). Performance pay, test scores, and student learning objectives. Economics of Education Review, 44, 114–125. doi:10.1016/j.econedurev.2014.11.002
Barge, D. J. (2013). Student learning objectives operations manual. Retrieved from http://www.gadoe.org/ School-Improvement/Teacher-and-Leader-Effectiveness/Documents/2013%20SLO%20Manual.pdf
Bergin, C. (2015). Using student achievement data to evaluate teachers. [White paper]. Retrieved from https://nee.missouri.edu/documents/NEE_White_PaperStudent_Achievement_in_TeacherEvaluation2015_5_20.pdf
Bracken, B. A. (2006). Bracken basic concept scale: Expressive (BBCS:E). San Antonio, TX: Pearson.
Briggs, D. C., Diaz-Bilello, E., Maul, A., Turner, M., & Bibilos, C. (2014). Denver ProComp Evaluation Report: 2010-2012. Boulder, CO: Colorado Assessment Design Research and Evaluation (CADRE) Center and the National Center for the Improvement of Educational Assessment. Retrieved on August 13, 2014, from http://www.colorado.edu/education/sites/default/files/attached-files/CADRE_NCIEA%20ProComp%20Full%20Report%20FINAL%202014.pdf
Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Cullimating findings from the MET project’s three-year study. MET Project Research Paper.
Center for Assessment. (2014). Using Baseline Data and Information to Set SLO Targets. Retrieved from http://www.ncbi.nlm.nih.gov/wp-content/uploads/7_U sing-Baseline-Data-and-Information_7.15.13.pdf
Chapman, L. H. (2014). The marketing of student learning objectives (SLOs): 1999-2014. Retrieved from http://vambooled.com/recommended-reading/
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. The American Economic Review, 104(4), 2633–2679. doi:10.1257/aer.104.9.2633
Community Training and Assistance Center. (2004). Catalyst for change: Pay for performance in Denver: Final report. Boston: Author. Retrieved from http://ctacusa.com/wp-content/uploads/2013/11/CatalystForChange.pdf
Community Training and Assistance Center. (2013). It’s more than the money. Boston: Author. Retrieved from http://ctacusa.com/wp-content/uploads/2013/11/MoreThanMoney.pdf
Crouse, K., Gitomer, D. H., & Joyce, J. (2016). An Analysis of the Meaning and Use of Student Learning Objectives. In Student Growth Measures in Policy and Practice (pp. 203–222). New York, NY: Springer.
Darling-Hammond, L., & Youngs, P. (2002). Defining “highly qualified teachers”: What does “scientifically-based research” actually tell us? Educational Researcher, 31(9), 13–25. doi:10.3102/0013189X031009013
de Jesus, N., & Lens, W. (2005). An integrated model for the study of teacher motivation. Applied Psychology, 54(1), 119–136. doi:10.1111/j.1464-0597.2005.00199.x
EngageNY. (2013). Critical Decisions within Student Learning Objectives (SLOs): Target Setting Models [PowerPoint slides]. Retrieved from http://www.enga
geo.ny.gov/sites/default/files/resource/attachments/critical_district_decisions_within_slos.pdf
Gegnon, D. J., Hall, E. L., & Marion, S. (2017). Teacher evaluation and local control in the US: An investigation into the degree of local control afforded to districts in defining evaluation procedures for teachers in non-tested subjects and grades. Assessment in
Education: Principles, Policy & Practice, 24(4), 489–505.

Georgia Department of Education. (2013). Georgia Department of Education SLO Operations Manual. Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/2013%20SLO%20Manual.pdf

Gill, B., Bruch, J., & Booker, K. (2013). Using alternative student growth measures for evaluating teacher performance: What the literature says. (REL 2013–002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Gill, B., English, B., Furgeson, J., & McCullough, M. (2014). Alternative student growth measures for teacher evaluation: Profiles of early-adopting districts. Mathematica Policy Research. Retrieved from http://fileseric.ed.gov/fulltext/ED544797.pdf

Goldhaber, D., & Walsh, J. (2011). Strategic pay reform: A students outcomes-based evaluation of Denver’s ProComp teacher pay initiative. Economics of Education Review, 31(6), 1067–1083. doi:10.1016/j.econedurev.2012.06.007

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher Value-Added at the High-School Level: Different Models, Different Answers?. Educational Evaluation and Policy Analysis, 35(2), 220–236. doi: 0162373712466938

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. American Educational Research Journal, 48(3), 794–831. doi:10.3102/0021962411409716

Joyce, J., Harrison, J. R., & Murphy, D. (2016). Evaluating Students with Disabilities and Their Teachers: Use of Student Learning Objectives. Teachers College Record, 118(14), 14.

Kimball, S. M., White, B., Milansowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe county. Peabody Journal of Education, 79(4), 54–78. doi:10.1207/s15327930pje7904_4

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.

Lamb, L. M., Schmitt, L. N. T., Gross, R., & Cornetto, K. M. (2013). Austin Independent School District (AISD) pilot teacher appraisal system update: 2012–2013 focus group and survey summary (DRE Publication No. 12.70). Austin, TX: Austin Independent School District. Retrieved from http://www.austinisd.org/sites/default/files/dre-reports/12.70_Austin_Independent_School_District_AISD_Pilot_Teacher_Appraisal_System_Update.pdf

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. Journal of Educational Measurement, 44(1), 47–67. doi:10.1111/j.1745-3984.2007.00026.x

Longchamp, J. C. (2017). The Effect Of Student Learning Objectives On Teachers And Teaching As Part Of The Teacher Evaluation Process: A Grounded Theory Study (Doctoral dissertation, The University of Vermont and State Agricultural College).

Makkonen, K. (2013). Exploring the Reliability and Validity of Pilot Teacher Ratings in a Large California School District (Doctoral dissertation, University of California, Berkeley).

Marion, S., Desrascole, C., Domaleski, C., Gong, B., & Díaz-Biello, E. (2012). Considerations for analyzing educators’ contributions to student learning in non-tested subjects and grades with a focus on Student Learning Objectives. Center for Assessment: Dover, NH. Retrieved on August 14, 2014, from http://www.ncee.org/publication_PDFS/Measurement%20Considerations%20on%20SLOs_052212.pdf

Marzano, R. J., & Toth, M. D. (2013). Teacher evaluation that makes a difference: A new model for teacher growth and student achievement. Alexandria, VA: ASCD.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Harell, L. S. (2003). Evaluating Value-Added Models for Teacher Accountability. Monograph. Santa Monica, CA: RAND Corporation.

Missouri Department of Elementary and Secondary Education. (2014). Student Learning Objectives Handbook. Retrieved from https://dese.mo.gov/sites/default/files/SLO-Handbook.pdf

Murphy, D. (2012). Where is the value in value-added modeling. Retrieved from https://images.personasessments.com/images/tmrs/Where_is_the_Value_in_Value_Added_Modeling.pdf 10.1094/PDIS-11-11-0999-PDN

New York State Education Department. (2013). Critical Decisions within Student Learning Objectives (SLOs): Target Setting Models [Video file]. Retrieved from https://www.engageNY.org/resource/critical-decisions-within-student-learning-objectives-slo-target-setting-models

Ohio Department of Education. (2013). A Guide to using SLOs as a locally-determined measure of student growth. Retrieved from http://education.ohio.gov/getattachment/Topics/Teaching/Educator-Evaluation-System/Ohio-s-Teacher-Evaluation-System/Student-Growth-Measures/Student-Learning-Objective-Examples/071513_SLO_Guidebook_FINAL.pdf.aspx

Paolini, A. (2011). Enhancing Teaching Effectiveness and Student Learning Outcomes. Journal of Effective Teaching, 15(1), 20–33.

Popay, J. P. (2011). Different tests, different answers the stability of teacher value-added estimates across countries. American Educational Research Journal, 48(1), 163–193. doi:10.3102/0021962411387577

Potenski, A. (2013). Flexibility for Fairness: Crafting Business Rules for Student Learning Objectives. Retrieved from https://gifrcenter.org/sites/default/files/docs/GTL_AskTeam_FlexForFairness.pdf

Prctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walsh, J. (2011). Making a difference in education reform: ProComp external evaluation report 2006–2010. Prepared for the Denver Public Schools. Denver: The Evaluation Center, University of Colorado.

Rice to the Top Technical Assistance. (2010). Measuring student growth for teachers in non-tested grades and subjects: A primer. Washington, DC: ICF International. Retrieved from http://nassauboces.org/cms/lib5/NY18000988/Centricity/Domain/156/NTS_PRIMER_FINAL.pdf

Raudenbush, S. W. (2004). HLM 6: Hierarchical linear and nonlinear modeling. Lincolnwood, IL: Scientific Software International, Inc. Retrieved from https://books.google.com/books?hl=en&id=VdmVzt6WtCOCGoI=fnd&pg=PA7&dq=Raudenbush+2004+&ots=kUXMQQ0q6d&sig=
Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. Education Finance and Policy, 4(4), 492–519. doi:10.1162/edfp.2009.4.4.492

Raudenbush, S. W. (2004). The impact of individual teachers on student achievement: Evidence from panel data. The American Economic Review, 94(2), 247–252. doi:10.1257/0002828041302244

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. Journal of Personnel Evaluation in Education, 8(3), 299–311. doi:10.1007/BF00973726

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24(3), 419–434. doi:10.1111/peps.1971.24.issue-3

Schmitt, L. N. (2011). AISD REACH program update, 2010–2011: Texas assessment of knowledge and skills: Growth and student learning objectives (DRE Publication No. 10.84). Austin, TX: Austin Independent School District. Retrieved from http://www.austinisd.org/sites/default/files/dre-reports/10.84_AISD_REACH_Growth_and_SLOs_2010-2011.pdf

Schmitt, L. N., & Lamb, L. M. (2014). AISD REACH Program Update, 2012–2013. Austin, TX: Austin Independent School District. Retrieved from http://www.austinisd.org/sites/default/files/dre-reports/12.96_AISDREACH_Program_Summary_of_Findings_2007_2008_Through_2012_2013_0.pdf

Schmitt, L. N. T., Cornetto, K. M., Lamb, L. M., & Imes, A. (2009). AISD REACH year 2 evaluation report I, 2008–2009 (DRE Publication No. 08.53). Austin, TX: Austin Independent School District. Retrieved from http://www.austinisd.org/sites/default/files/dre-reports/08.53_AISD_REACH_Year_2_Evaluation_Report_I_2008-2009.pdf

Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. Sociological Methods & Research, 22(3), 342–363. doi:10.1177/0049124194022003004

Trochim, W. M., & Donnelly, J. P. (2001). Research methods knowledge base. Retrieved from http://www.anatomyfacts.com/research/researchmethodsknowledgebase.pdf

Utah State Office of Education. (2014). Utah Guidance and Toolkit for Student Learning Objectives. Retrieved from http://www.schools.utah.gov/CURR/educatoreffectiveness/Student-Growth/SLOToolkit.aspx.