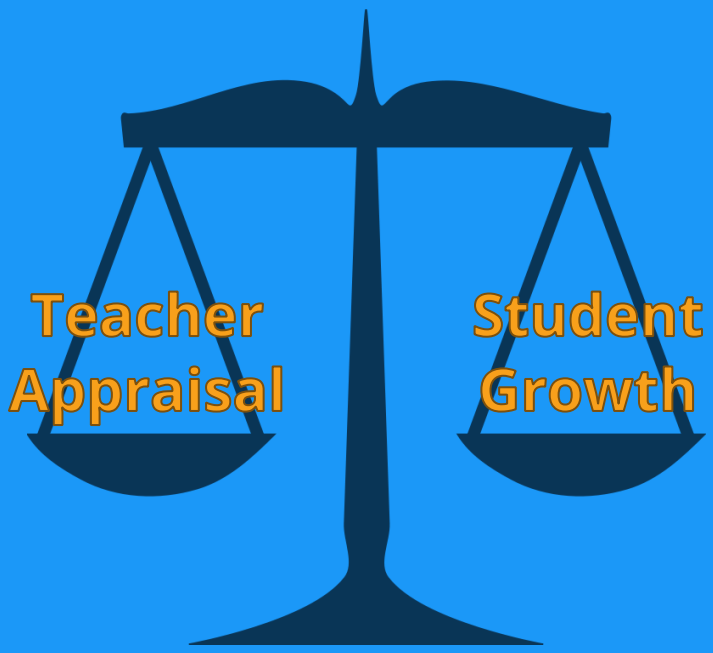
November 2015
Publication Number 14.85

Student Learning Objectives (SLOs)

Analysis of Student Growth in 2013–2014, by Type and Source of Assessment



Department of Research and Evaluation
AUSTIN INDEPENDENT SCHOOL DISTRICT

AISD Guide for Developing Student Learning Objectives

Needs Assessment / Rational
What are the needs?

Learning Content / Context and Student Group
What and who is targeted?

Learning Objective
What will students learn?

Outcome Assessment
How will you know whether they learned it?

Student Growth Target
What is your goal for student achievement?

Executive Summary

This report examines the practical implications of student learning objective (SLO) assessment decisions for teacher appraisal. We present student-level growth data from SLO assessments administered in 2013–2014. Growth data are compared according to type (i.e., multiple-choice or rubric/performance-based) and source (i.e., common or teacher-created) of assessment. The work was a follow up to previous SLO research (Schmitt, 2014) that examined 2013–2014 teacher-level SLO data. Results support previous findings and suggest additional key findings. Overall, results support teachers' use of the Austin Independent School District (AISD) common SLO assessments.

Student growth was comparable for common assessments and teacher-created rubric/ performance-based assessments, but was lower for teacher-created multiple-choice assessments.

Measured student growth did not differ between common multiple-choice, common rubric/performance-based, and teacher-created rubric/performance-based assessments. However, student growth on teacher-created multiple-choice assessments was significantly worse than growth measured with other assessments.

Students were least likely to meet growth targets on teacher-created multiple-choice assessments.

The percentage of students who met growth targets on teacher-created multiple-choice assessments (60.3%) was less than that for common multiple-choice (75.6%), common rubric/performance-based (81.5%), and teacher-created rubric/performance-based assessments (81.4%).

Students measured with teacher-created multiple-choice assessments were more likely than other students to show decline from the beginning to the end of year.

The percentage of students showing negative growth was highest for teacher-created multiple-choice assessments (11.0%). In comparison, 7.6% of students assessed with common multiple-choice assessments, 7.5% of students assessed with common rubric/ performance-based assessments, and 3.2% of students assessed with teacher-created rubric/performance-based assessments showed negative growth from the beginning to the end of the school year.

The SLO work with percentage-based measures highlighted assumptions about the comparability of growth percentages.

Transforming raw scale pretest-posttest differences into percentages for the purpose of standardizing and comparing student growth can be common practice, given different types of assessments. However, the actual change in scale distance from pretest to posttest can vary considerably for the same percentage of growth. Results suggest the need to question whether growth percentages are truly comparable.

Table of Contents

List of Figures

List of Tables

List of Insets

AISD Guide for Developing Student Learning Objectives

Needs Assessment / Rational
What are the needs?

Learning Content / Context and Student Group
What and who is targeted?

Learning Objective
What will students learn?

Outcome Assessment
How will you know whether they learned it?

Student Growth Target
What is your goal for student achievement?

Introduction

Because teachers in Austin Independent School District (AISD) are given the choice to create their own student learning objectives (SLO) assessments or choose from a bank of district preapproved assessments, the aim of the SLO growth analysis was to examine the practical implications of SLO assessment decisions for teacher appraisal. The percentage of students meeting SLO growth targets contributes to teachers' appraisal scores. Thus, to ensure equity in teacher appraisal, different SLO assessments of otherwise similar students should fairly represent the degree of student growth. This study explores whether specific assessment choices may advantage or disadvantage teachers.

Key Findings

Teachers should be cautious about creating their own multiple-choice SLO assessments. Teachers who choose to create their own multiple-choice assessments may earn fewer appraisal points than do other teachers.

Overall, results support teachers' use of the AISD common SLO assessments. Measured student growth did not differ for common multiple-choice, common rubric/performance-based, and teacher-created rubric/performance-based assessments. However, student growth on teacher-created multiple-choice assessments was significantly worse than on other assessments, as were the percentages of students who met growth targets. Additionally, scores on teacher-created multiple-choice assessments were more likely than scores on other assessments to decline from the beginning to end of year. The differences in student growth between assessments translated to about 13 appraisal points, on average. This represented 22% of the 60 possible SLO appraisal points and 3% of the 400 possible total appraisal points.

Prior SLO Research in AISD

This study is the first to examine student-level performance on SLO assessments in AISD. Prior studies of SLO performance examined teacher-level SLO performance data across various combinations of school, subject, level, teaching area, assessment source, and assessment type.[1] The most recent AISD SLO study noted differences in teacher-level SLO performance according to SLO assessment type (i.e., multiple-choice or rubric/performance-based), and recommended further examination of the thresholds for meeting SLO growth targets (Schmitt, 2014).

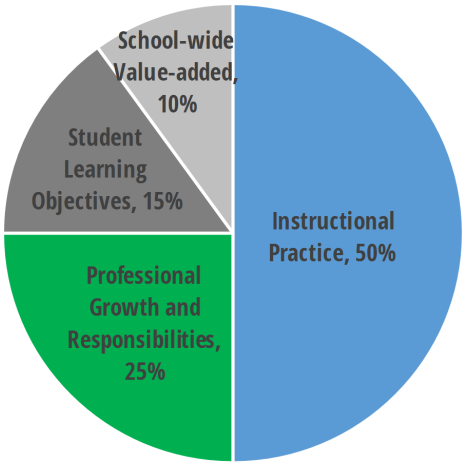
[1] See for example:
Schmitt (2014). AISD REACH Program Update: Student Learning Objective Assessments (DRE Publication No. 13.89 RB).
Schmitt, Lamb, Cornetto, & Courtemanche (2014). AISD REACH program update, 2012–2013: Student learning objectives (DRE Publication No. 12.83b).
Schmitt (2011). AISD REACH program update, 2010–2011: Texas Assessment of Knowledge and Skills growth and student learning objectives (DRE Publication No. 10.84 RB).

Implications of Growth on Appraisal Points Earned

Under AISD's appraisal system, Professional Pathways for Teachers (PPfT),[2] SLOs contribute 15% to a teacher's overall appraisal points earned. Specifically, with the PPfT appraisal system (Figure 1), teachers earn appraisal points from four separate components, a single measure of collective attribution and three measures of individual attribution. The four components are:

- School-wide value-added rating (10%)
- Individual SLOs (15%)
- Professional growth and responsibilities (25%)
- Instructional practice (50%)

Figure 1.
PPfT Appraisal Overview



The maximum appraisal points possible are 400: 40 points from school-wide value-added, 60 points from SLOs, 100 points from professional growth and responsibilities, and 200 points from instructional practice. SLO appraisal points are computed using the following formula: (% met growth target / 25) x 15. Thus, a teacher with 100% of his or her students meeting the growth target would earn 60 SLO appraisal points and a teacher with 25% of his or her students meeting the growth target would earn 15 SLO appraisal points. A teacher's total appraisal points earns him or her one of five final rating categories under PPfT (Table 1).

Table 1.
Final Rating Categories, by Total Appraisal Points Earned under PPfT

Total Points Earned	Final Rating
85 <= Points < 200	Ineffective
200 <= Points < 257	Minimally Effective
257 <= Points < 314	Effective
314 <= Points < 370	Highly Effective
370 <= Points <= 400	Distinguished

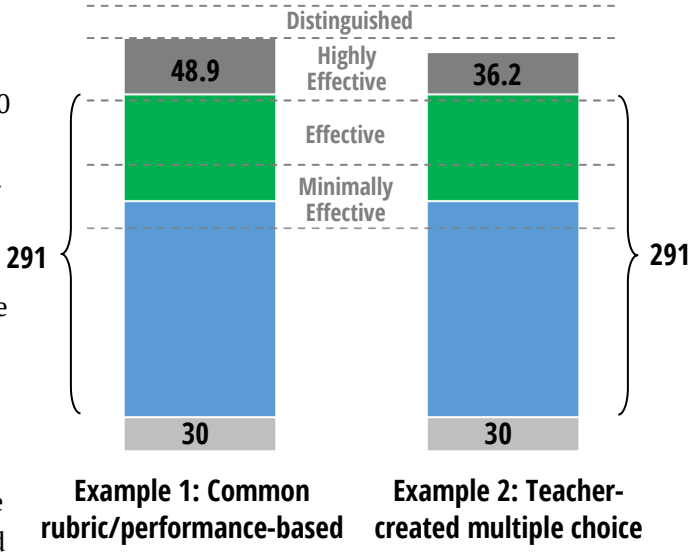
The present study examines the percentage of students who met teachers' growth targets in 2013–2014 according to the category of SLO assessment used. Using the mean percentage of students meeting growth targets per assessment category (discussed later, see for example Figures 8 and 9), the average points earned from SLOs can be estimated at 45.4 for teachers using common multiple-choice assessments, 48.9 with common rubric/performance-based assessments, 36.2 with teacher-created multiple-choice assessments, and 48.8 with teacher-created rubric/performance-based assessments. The 13 point gap alters the total point ceiling available to otherwise similar teachers within the same school who potentially only differ by choice in SLO assessment. The gap in appraisal scores is more dramatic when comparisons are made across campuses with different school-wide value-added scores (i.e., a measure of collective attribution).

The following is an example case of similar teachers only differentiated by choice of SLO assessment. The average teacher using a common rubric/performance-based assessment at a campus receiving a school-wide value-added rating of 3 (i.e., 30 appraisal points) would

[2] See the AISD PPfT Support Guide for more details http://www.austinisd.org/sites/default/files/dept/reach/PPFT_Support_Guide_Final_15-16.pdf

need to earn about 291 out of 300 possible points on the instructional practice and professional growth and responsibilities components of PPfT to hit the threshold for receiving a final rating of distinguished (Example 1 in Figure 2). The average teacher at the same campus also earning 291 out of 300 possible points for instructional practice and professional growth and responsibilities but using a teacher-created multiple-choice assessment would receive a final rating of highly effective (Example 2 in Figure 2). In fact, a distinguished rating would not be achievable for a teacher with the average SLO appraisal points for teacher-created multiple-choice assessments at a campus with a school-wide value-added score of 3 (i.e., 36.2 + 30 + 300 = 366.2).

Figure 2.
Example Differentiation by SLO Assessment Only



The average teacher using a teacher-created multiple-choice assessment at a campus receiving a school-wide value-added (V-A) rating of 4 (i.e., 40 appraisal points) would need to earn about 294 out of 300 possible points on the instructional practice and professional growth and responsibilities components of PPfT to receive a final rating of distinguished. However, a score of only 281 out of 300 would be necessary for the average teacher using a common rubric/performance-based assessment at the same school to earn a final rating of distinguished (Figure 3A).

When collective attribution (i.e., school-wide value-added) is factored into the example case, the altered total point ceiling available becomes more apparent. Figure 3A shows how the average teacher using a teacher-created multiple-choice assessment at a campus receiving a school-wide value-added rating of 4 could receive a final rating of distinguished. However, similar teachers with the average SLO appraisal points for teacher-created multiple-choice assessments at campuses with lower school-wide value-added ratings could not earn a rating of distinguished even if earning all 300 points on the instructional practice and professional growth and responsibilities components (Figure 3B).

Figure 3A.
Example Differentiation by All Individual Attribution Measures

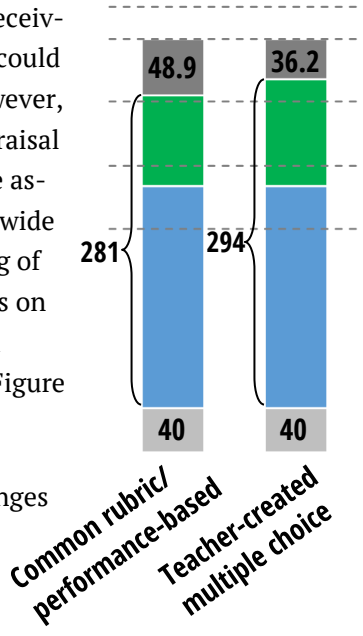
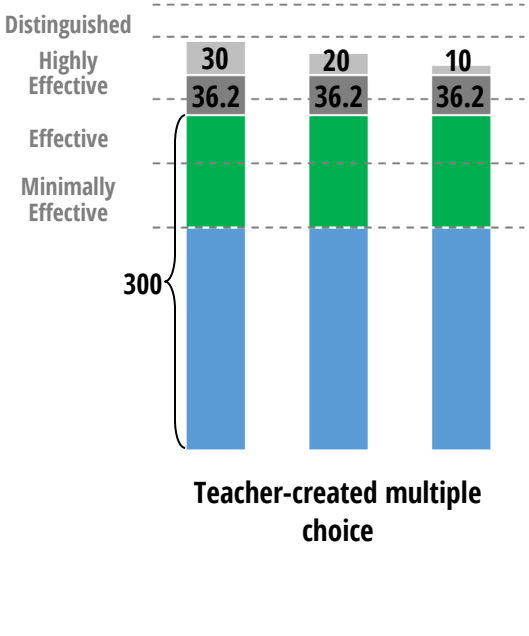
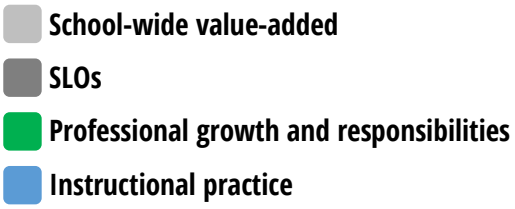


Figure 3B.
Example Differentiation by Collective Attribution Only

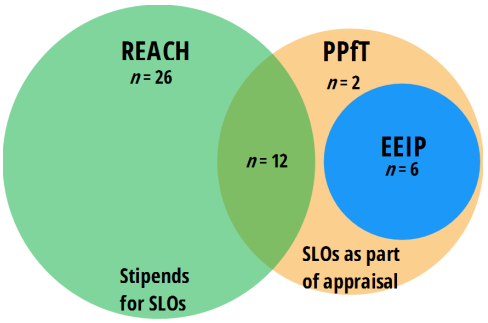




The following sections describe the challenges associated with the variety of SLO assessment scales used, propose a method for establishing a common way to examine student growth for different assessments, and demonstrate the relative amount of student growth achieved according to assessment category.

Inset 1.

A Timeline of Student Learning Objectives (SLOs) in AISD



REACH PPfT EEIP

2007

Year 1: 2007–2008

SLOs introduced to 9 campuses as part of a strategic compensation program, AISD REACH. Participants included teachers, principals, and some instructional coaches.

- Two individual teacher-created SLOs per teacher; one for the entire class, the other could target a subgroup
- Campus support for the SLO process provided from the AISD REACH central office SLO team

2008

Year 2: 2008–2009

Two campuses were added to AISD REACH in Year 2 bringing the total to 11 campuses. Participants also expanded to include librarians, assistant principals, and all instructional coaches.

- One of two SLO assessments must be from new assessment bank for core areas
- Tiered student targets allowed

2009

Year 3: 2009–2010

AISD REACH expanded to 15 campuses in Year 3. Participants expanded to include counselors.

- Common assessment bank for SLOs abandoned
- Campus-based SLO facilitators introduced on some REACH campuses as an additional method of support for the SLO process

2010

Year 4: 2010–2011

AISD REACH expanded to 19 campuses in Year 4.

- Team SLOs implemented
- One individual and one team or additional individual SLO

2011

Year 5: 2011–2012

AISD REACH expanded to 28 campuses in Year 5.

- Half the distance to a perfect score growth target introduced
- From 2011–2012 to 2013–2014, for the first time, SLOs were also part of a pilot teacher appraisal system at some REACH schools

2012

Year 6: 2012–2013

AISD REACH expanded to 38 campuses in Year 6.

- New district-created, preapproved assessment bank introduced

2013

Year 7: 2013–2014

The six cohorts of AISD REACH continued into a second and third year as a 38 campus district sample in Years 7 and 8.

- Preapproved assessment bank expanded

2014

Year 8: 2014–2015

In the final year of AISD REACH, SLOs were also used in AISD's new appraisal system, Professional Pathways for Teachers (PPfT), on 20 campuses (both REACH and non-REACH). Six of the 20 PPfT campuses participated in AISD's Educator Excellence Innovation Program (EEIP).

- One individual SLO per teacher for EEIP and PPfT
- Preapproved assessment bank expanded

2015

Year 9: 2015–2016

AISD will continue to use SLOs in 2015–2016 for EEIP and PPfT (with expansion from 20 to 35 campuses).

2016

Year 10 and beyond: 2016–2017+

Use of SLOs is anticipated to carry forward as part of the PPfT in AISD.

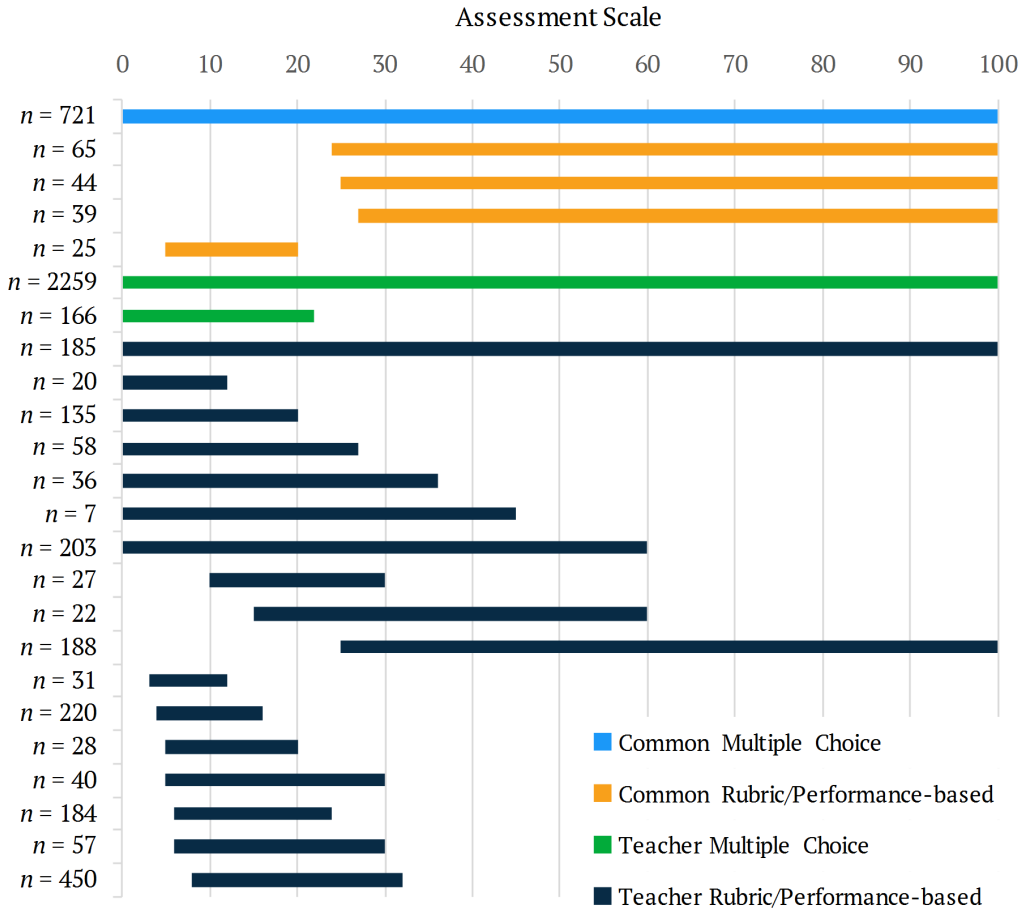
2017+

Assessment Scales Used

In AISD, teachers could choose either to create their own SLO assessments or to select from a bank of common district-approved and vetted SLO assessments. Furthermore, teachers had the option of using a multiple-choice or rubric/performance-based assessment type. Of the 5,210 middle and high school students sampled from classrooms of 154 educators, 721 were assessed with common multiple-choice assessments, 173 were assessed with common rubric/performance-based assessments, 2,425 were assessed with teacher-created multiple-choice assessments, and 1,891 were assessed with teacher-created rubric/performance-based assessments.

Across the four assessment categories, 24 different assessment scales were used in the sample. Figure 4 shows the different scales used for each assessment source and type combination. Within common multiple-choice assessments only one scale was used (although the point increments per question varied with the number of questions on the assessment). Within common rubric/performance-based assessments four different scales were used. Within teacher-created multiple-choice assessments two different assessment scales were used. Within teacher-created rubric/performance-based assessments 17 different assessment scales were used.

Figure 4.
Assessments varied considerably in the scales employed.



Source. REACH SLO database using the following criteria: SLO = individual SLO 1; level = high school or middle school; content = reading, writing, or math.
Note. Counts are students per assessment.

Scale Diversity: Fair versus Equal?

"Equal" means exactly the same or exactly alike. **"Fair,"** however, allows room for just dissimilarities that result in equitable although not equal circumstances. In education, if the desired outcome is to be equal, then it is typical to require that process have the needed flexibility to be individually fair and equitable.

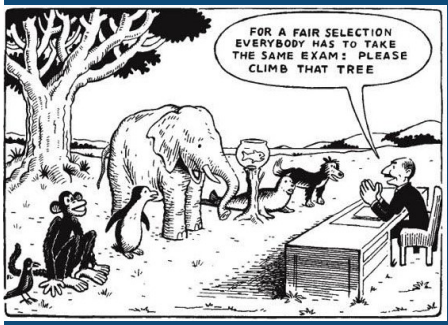
Teacher-created rubric/performance-based assessments had the most variability in scales. Given the common belief in education that treating students fairly does not equal treating them the same, does the same apply to teachers' appraisals, given the diversity of subjects and levels taught?

The question from an assessment perspective is:

... To fairly assess students across subjects and levels, do the assessments themselves have to be equal?

The question from an appraisal perspective is:

... To fairly judge student growth as part of teacher appraisal, do the assessments have to be equal?



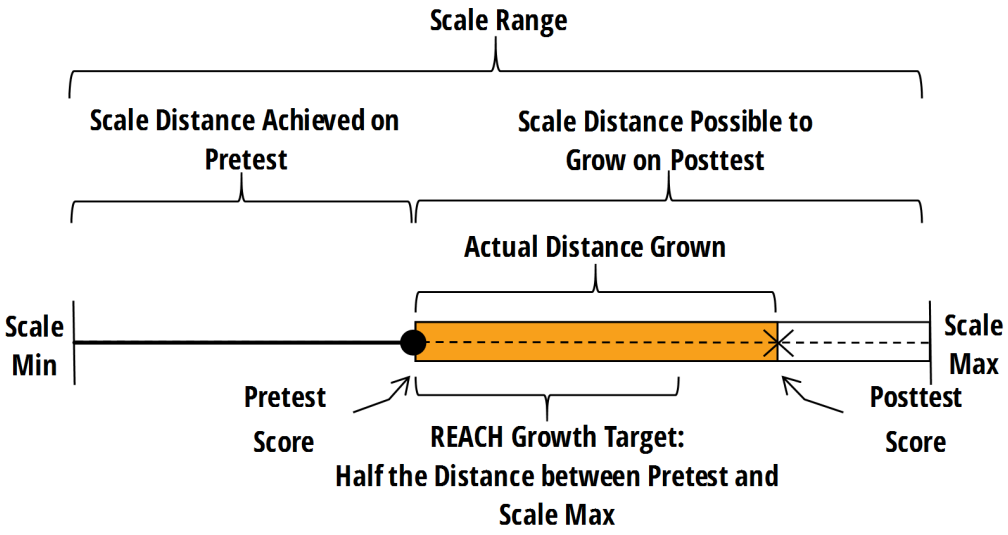
SLO Measures

Due to the diversity of scales used for SLO assessment, we needed a common way to look at pretest performance and student growth. Two percent of scale-based measures were computed in response as extensions of the REACH concept for setting growth targets: a measure of pretest performance as the percentage of the assessment scale achieved and a measure of student growth as the percentage of the scale distance possible to grow on the posttest.

REACH Growth Target

The policy for setting student SLO growth targets followed a uniform growth target formula for half the distance to a perfect score, meaning that regardless of assessment scale or pretest performance, growth targets were at a minimum half of the distance between the student's pretest score and the highest possible score on the scale (i.e., [max-pretest] x 0.5). The result was a growth target value, in raw scale units, equal to 50% of the distance available to grow on the assessment's scale. Figure 5 visually disaggregates the components of student growth and highlights the REACH growth target calculation.

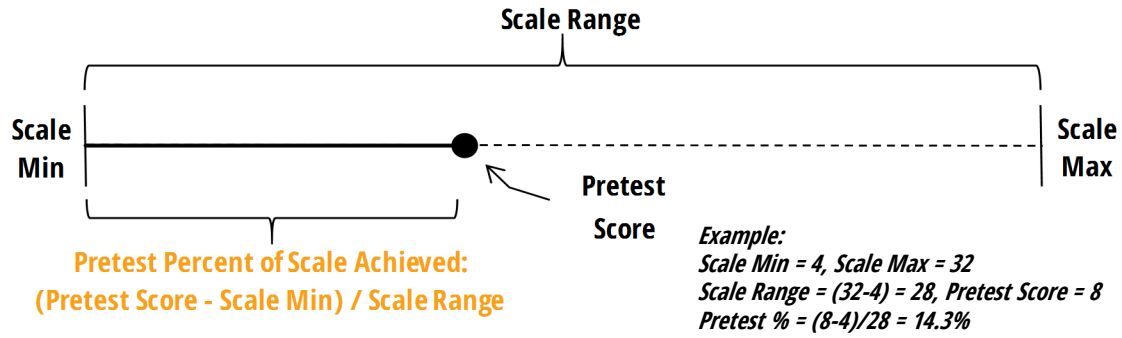
Figure 5.
The Components of Student Learning Objectives Growth Measurement: Scale Range, Pretest Scale Distance Achieved, Scale Distance Possible to Grow, Pretest Score, Posttest Score, Scale Minimum/Maximum, and Posttest Actual Distance Grown



Pretest Performance

Due to the variety of possible score ranges (e.g., 4–16, 0–100, 8–32), pretest performance was not comparable without some transformation. To obtain comparable scores, we subtracted the minimum scale value from the pretest score and then divided the difference by the range of the scale (i.e., [pretest-min]/scale range). The resulting percentage-based measure of pretest performance provided a common way of thinking about pretest performance across assessment scales. Figure 6 shows the components of pretest performance.

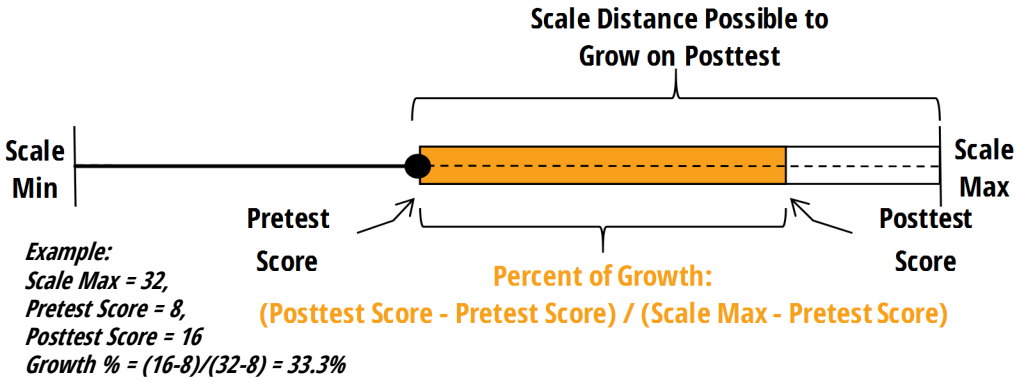
Figure 6.
Computed Measure of Pretest Performance (i.e., Pretest Percentage of Scale Achieved)



Student Growth

Where pretest performance represented the relative scale distance achieved on the pretest, the student growth measure represented the pretest-posttest difference relative to the remaining scale distance possible to grow. Student growth was computed by dividing actual growth (i.e. posttest score-pretest score) by the assessment scale distance possible to have grown on the posttest (i.e., scale max-pretest score), resulting in a measure of actual growth as a percentage of the distance possible to grow.[3] The resulting percent-based measure of student growth provided a common way of thinking about growth across assessment scales, where the larger the transformed percentage, the more growth measured relative to the maximum amount of growth possible. Figure 7 shows the components of student growth.

Figure 7.
Computed Measure of Student Growth (i.e., Percentage of Growth Possible)



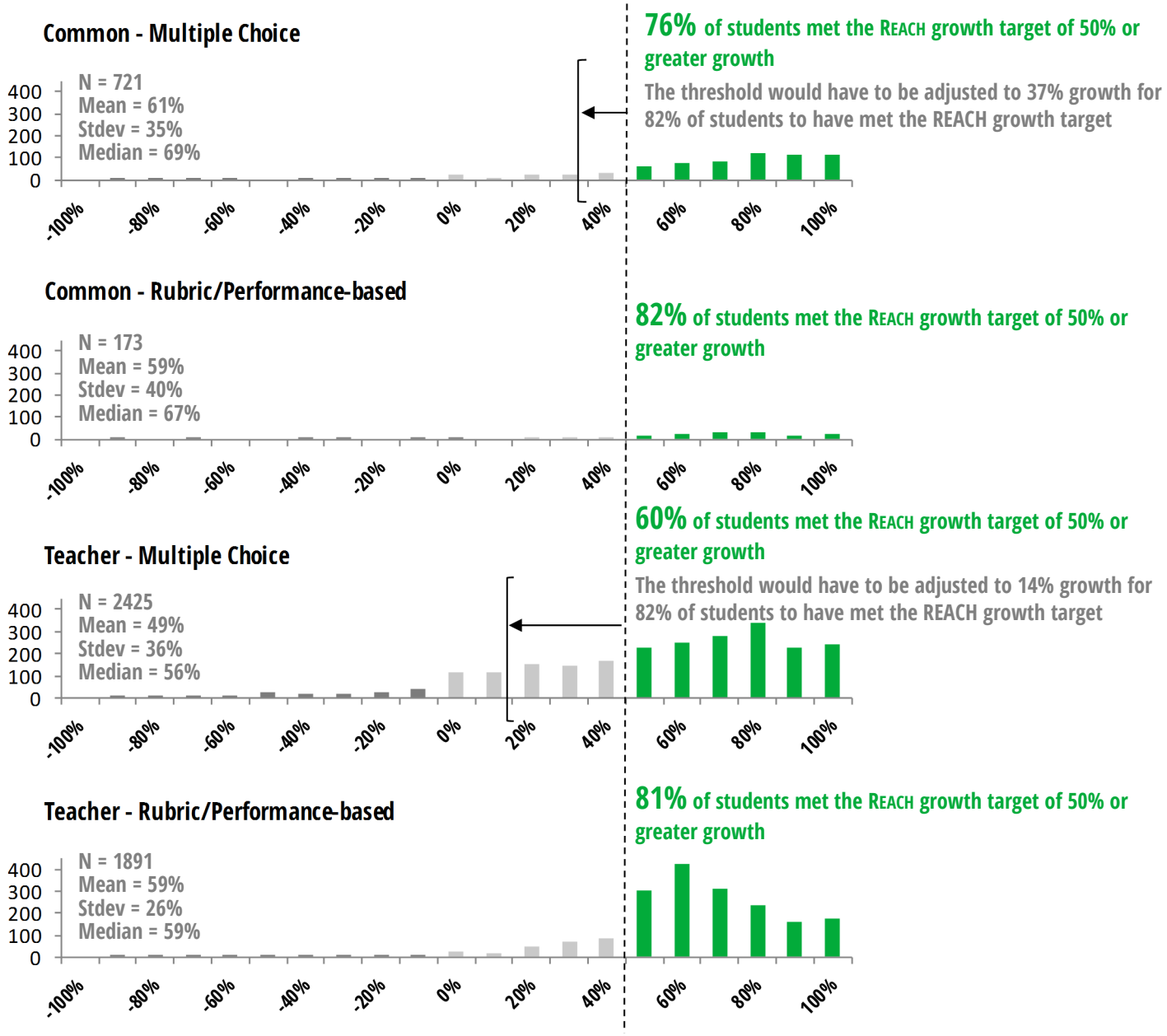
[3] For students scoring the maximum points possible on the pretest, no growth percentage possible outcome measure could be computed (i.e., one cannot divide actual growth by zero); consequently, *n*'s may be less on this outcome measure than represented in the sample.

Analysis of Student Growth Distributions

Among the four assessment categories, student growth was lowest for teacher-created multiple-choice assessments (49% of possible growth, on average, from pre- to post-assessment). In fact, the average growth students demonstrated on teacher-created multiple-choice assessments was less than the criterion for acceptable growth, half of the distance between pretest and a perfect score (Figure 8).

Figure 8.

Fewer students met the REACH growth target with multiple-choice assessments than with rubric/performance-based assessments.



Source. REACH SLO database using the following criteria: SLO = individual SLO 1; level = high school or middle school; content = reading, writing, or math.

The student growth descriptive statistics for the four assessment categories are shown in Table 2 disaggregated by assessment source and assessment type.

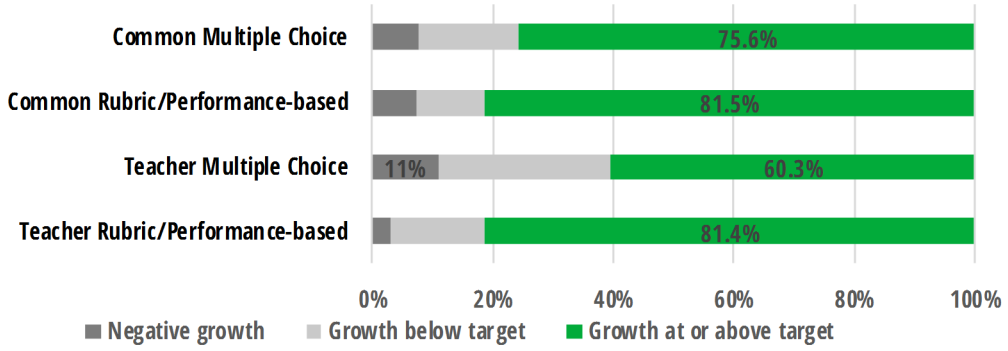
Table 2.
Student Growth Descriptive Statistics

Source	Type	N	Mean	Stdev	Median	Mode	Min	Max
Common	Multiple-choice	721	61%	35%	69%	80%	-90%	100%
	Rubric/ Performance-based	173	59%	40%	67%	80%	-90%	100%
Teacher	Multiple-choice	2425	49%	36%	56%	80%	-90%	100%
	Rubric/ Performance-based	1891	59%	26%	59%	60%	-90%	100%

Note. Modes were based on binned values rather than raw values. Binning involved rounding to one decimal place. Raw modes were 1.0, 1.0, 0.5, and 0.5, respectively.

A closer look at the student growth distributions for each assessment category revealed that students were least likely to meet the growth criterion and also were most likely to demonstrate negative growth (i.e., to score lower on the post-assessment than on the pre-assessment) on teacher-created multiple-choice SLO assessments. Figure 9 shows the percentages of students who met the acceptable growth target, demonstrated growth less than the target, or had negative growth for each assessment category.

Figure 9.
Teacher-created multiple-choice SLO assessments showed the lowest percentage of students meeting growth targets and the highest percentage of students demonstrating negative growth.



Source. REACH SLO database using the following criteria: SLO = individual SLO 1; level = high school or middle school; content = reading, writing, or math.

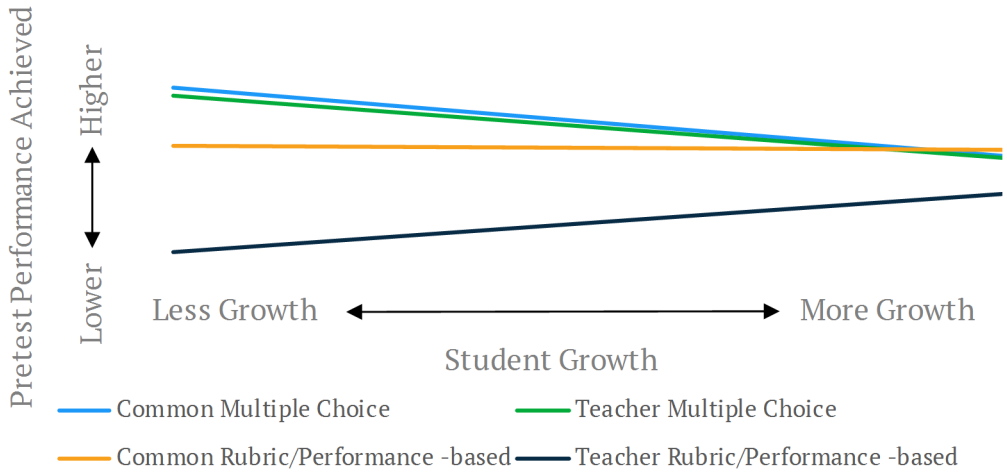
Differences in Student Growth, by Assessment Characteristics

When examining student growth, the students' starting point can matter. A weak, yet significant correlation was found between student growth and pretest performance. Students who scored higher on the pretest demonstrated less growth than did students who scored lower on the pretest. On average, every percentage point higher students scored on the SLO pretest measure was associated with a 1.7 point decrease in measured student growth. However, analysis of pretest performance by assessment category suggested pretest scores were not comparable across assessment categories[4] (Figure 10). Thus, we needed to control for the difference in pretest performance and the relationship between pretest scores and growth in the analysis of differences in student growth across assessment categories.

Overall, pretest performance was associated with growth, and pretest performance differed between assessment categories. Therefore, we accounted for mean pretest performance in comparisons of student growth.

Figure 10.

Pretest performance and the relationship of pretest scores to student growth differed across assessment groups.



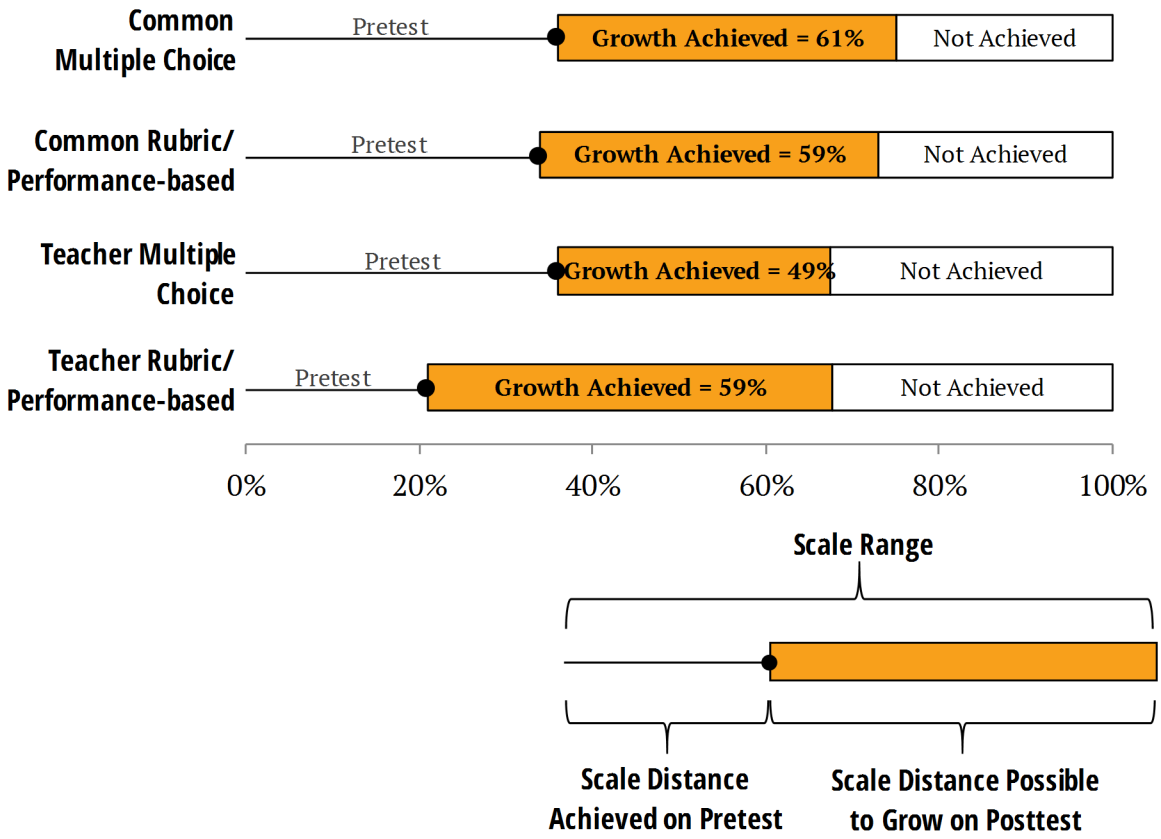
Source. REACH SLO database using the following criteria: SLO = individual SLO 1; level = high school or middle school; content = reading, writing, or math.

[4] The omnibus linear model on pretest scores indicated differences between groups. Standardized mean differences ranged from as small as 0.01 to as large as 0.8; the mean across pair-wise differences was 0.41. Although a non-experimental design, the Institute of Educational Sciences recommends examining baseline equivalence for quasi–experimental designs, and if groups differ by more than 0.05 but less than 0.25, then statistical control is needed to account for baseline differences; if groups differ by more than 0.25, baseline equivalence cannot be assumed.

Analyses were run with and without a pretest score covariate.[5] After statistically controlling for pretest performance, the mean for student growth remained significantly lower for teacher-created multiple-choice assessments than for other assessments. Figure 11 shows mean student growth from mean pretest performance for each assessment group. Common multiple-choice, common rubric/performance-based, and teacher rubric/performance-based assessments did not differ from each other; all three groups demonstrated about 60% growth. Teacher-created multiple-choice assessments differed from the three other assessment groups, showing significantly less growth.[6]

Figure 11.

Assessment groups differed in mean measured student growth, but also differed in mean pretest performance.



Source. REACH SLO database using the following criteria: SLO = individual SLO 1; level = high school or middle school; content = reading, writing, or math.

Note. The model for assessment source by assessment type was significant [$F(3, 5206) = 42.8$, $p < 0.0001$, $R^2 = 0.02$] and showed a significant interaction of source x type ($p < 0.0001$) as well as significance of both main effects. Although both models showed poor fit, controlling for the effect of pretest scores improved the fit of the overall model ($R^2 = 0.04$, +0.02 from basic interaction model). The ordinal pattern of group means and interaction of source and type did not change when controlling for pretest scores.

[5] Analysis of covariance with pretest performance as the covariate was used to statistically control for pretest scores in an analysis of differences in student growth between assessment groups.

[6] Pair-wise comparisons of assessment groups were performed using Fisher's least significant difference. Differences between assessment combinations were evaluated using $\alpha = .05$.

Is Growth Equitable? Some Issues to Consider

Analyzing SLO data with percentage-based measures was necessary to compare different assessments on a common scale. However, percentage-based measures remove the different scale-dependent properties that exist in raw-scale units. Differences between scale properties in raw-scale units and percentage units are important to consider.

Scale Range

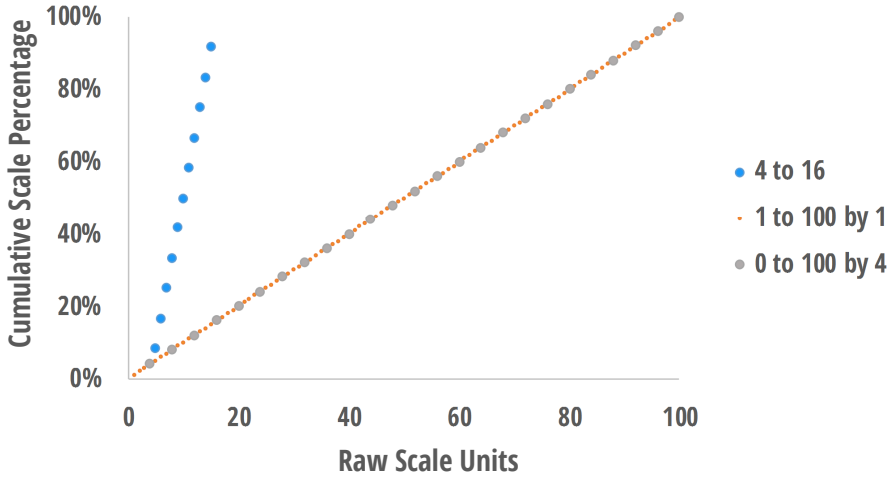
Scale range can be defined as the difference between the maximum and minimum scale values. The size of the assessment scale range and number of increments have a dramatic effect on the amount of change, relative to the entire scale, that is captured in a one-unit increment on the scale. For example:

- On a rubric scored 4–16 in 12 one-point increments, a one-unit increment equals 8.3% of the scale.
- On a rubric scored 8–32 in 24 one-point increments, a one-unit increment equals 4.2% of the scale.
- On a rubric scored 0–16 in 16 one-point increments, a one-unit increment equals 6.25% of the scale.
- On an assessment scored 0–100 in 100 one-point increments, a one-unit increment equals 1% of the scale.
- On an assessment scored 0–100 in 25 four-point increments, a one-unit increment equals 4% of the scale.

Should it be more or less difficult to move one unit on small scales or large scales? If the percentage increments of change on a 4 to 16 scale are 8.3% per score unit (i.e., scale range = 12, 8.3% x 12 = 100%) but are 1% on a 0 to 100 scale, are the amounts of measured learning proportional? In other words, is a greater amount of learning reflected in the former scale's one-unit increments? Figure 12 shows the difference in cumulative increments of raw-scale units from the assessment scale minimum to maximum. One question matters more on the 4 to 16 scale than the 0 to 100 scale in four-point increments; one question matters the least in the 0 to 100 scale in one-point increments. Students measured on the 4 to 16 scale would likely have less dispersion of scores and more students with the same score than would students measured on the 0 to 100 scale.

Figure 12.

Cumulative Percentages of the Assessment Scale for Each One-Unit Change in the Raw Scale



Scale Distance Possible to Grow

Two issues related to the comparability of growth percentages are highlighted: comparability across different assessment scales and comparability within assessment scales. Both issues are related to scale distance available to grow.

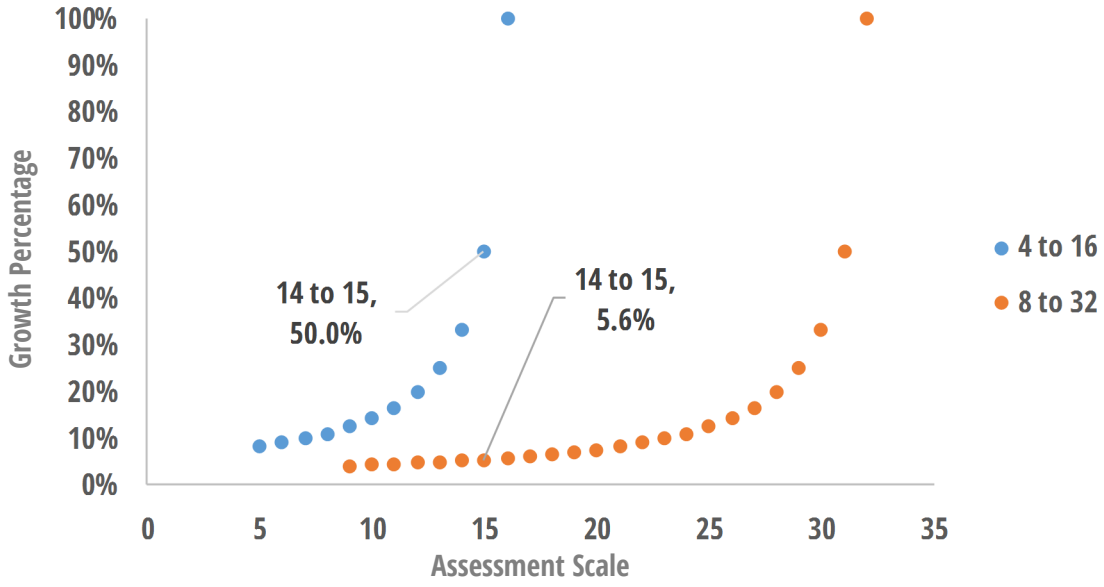
The percentage of growth achieved for a one-unit increment on an assessment scale depends upon the scale available to grow, and the scale available to grow can vary across assessments. For example, on a rubric scored 4 to 16 compared with a rubric scored 8 to 32, growth for a one-unit change increases unequally and nonlinearly throughout the same range of raw units because the scale available to grow differs. The growth associated with a pretest-to-posttest change in scores (for 4–16 versus 8–32, respectively) is:

- 10 to 11 = 16.67% vs. 4.55%
- 11 to 12 = 20% vs. 4.76%
- 12 to 13 = 25% vs. 5%
- 13 to 14 = 33.3% vs. 5.26%
- 14 to 15 = 50% vs. 5.56%
- 15 to 16 = 100% vs. 5.88%

Are different amounts of learning captured (e.g., 50% vs. 5.56% for a pretest score of 14 and posttest score of 15 on rubrics scored 4–16 and 8–32, respectively) or just differences in the properties of the scale? Figure 13 shows the distribution of growth increments in one-unit changes of the raw scale listed in the bullets above.

Figure 13.

Example Percentages of Growth for Each One-Unit Change in the Raw Scale on 4 to 16 and 8 to 32 Scales



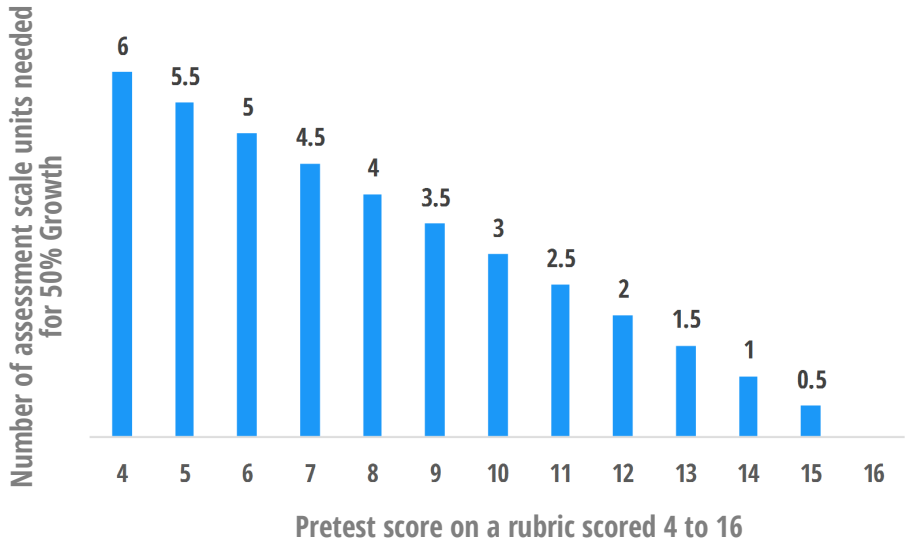
The scale available to grow can also vary within an assessment depending on pretest score. The pretest dependency consequence is twofold on a single assessment: (a) a constant number of assessment scale increments results in different percentages of growth and (b) movement to a common growth target requires different numbers of assessment scale increments. The current discussion focuses on the latter, given the REACH growth target context (i.e., 50% growth or half of the distance between the pretest score and the highest

possible score on the scale). The REACH growth target was designed to accommodate different growth expectations, in raw-scale units, for students scoring differently on the pretest, while maintaining a consistent 50% or greater growth-percentage expectation for all students regardless of pretest. For example, on a rubric scored 4 to 16 (see Figure 14):

- Pretesting at 6 requires 5 scale units for 50% growth.
- Pretesting at 7 requires 4.5 scale units for 50% growth.
- Pretesting at 8 requires 4 scale units for 50% growth.
- Pretesting at 9 requires 3.5 scale units for 50% growth.
- Pretesting at 10 requires 3 scale units for 50% growth.

Figure 14.

Example of the Varying Numbers of Raw Assessment Scale Units Needed for 50% Growth, Given Different Pretest Scores on a 4–16 Assessment Scale



Given the potential issues of comparability between growth percentages computed as a percentage of the scale available to grow, should percentage growth be considered equal? In other words, is 50% growth always equal? The nature of this conversation about comparability of growth exists at multiple levels of granularity, within which the question at hand is reframed to *when is growth comparability important*?

When is Comparability Important?

The difference between the *whether* and the *when* of growth percentage comparability is nontrivial. SLOs are intended to be teachers' and practitioners' tool for assessing the pretest-to-posttest growth of their own students. The granularity of the conversation may be framed as comparability within the classroom, at the grade/school/district level, and at the state level. A more complex and robust conversation about scale-independent measures of student growth at the state level exists in the literature on student growth percentiles (e.g., Betebenner, 2008). However, the percentile conversation extends beyond practitioners' classroom measure of *their own students'* pretest-to-posttest change and into measures of student growth among all "academic peers" in the state.

Consideration of whether growth percentages are comparable is certainly relevant when the comparison is academic peers in the state; however, a state assessment is far from the

practitioner's classroom assessment intended for SLOs. When growth percentages are bound to a single teacher's students (presumably measured with the same SLO assessment), the issue of whether growth percentages are comparable is limited to contrasts between large raw scale increases for students scoring low on the pretest and small raw scale increases for students scoring high on the pretest.

For example, Figure 14 demonstrated that on a 4 to 16 scale, a student scoring a 4 on the pretest would have to grow 6 scale points to achieve 50% growth, in contrast with a student scoring 14 on the pretest only needing to grow one scale point. Peer grouping on baseline data within the classroom is one method of accounting for the different starting points, either by focusing entirely on a subset of students and a common growth goal (an available strategy for secondary AISD REACH SLOs[7]) or by establishing tiered student-performance groups and associated growth targets in a single growth goal (an available strategy for AISD REACH SLOs and in other district appraisal systems, e.g., Jefferson County Public Schools[8]).

The comparability of growth percentages is complicated at the grade, school, and district level by the inclusion of different types of SLO assessments and aggregation of the measured growth across content areas. The complication is further exacerbated by the narrow subsets of skills assessed with SLOs when they are aggregated across content areas and grade levels. When growth percentages calculated from classroom-based SLOs are aggregated beyond the classroom to the school or district level, all starting points, all assessment scales, and all grade- and content-specific skills are held equal. The conversation about comparability of growth percentages at the intermediate level (e.g., grade, school, and district) is elevated when the practitioner measures are used as part of teacher appraisal (i.e., high-stakes testing), a purpose increasingly associated with measuring student growth at the at the grade, school, and district levels.

[7] See the AISD REACH SLO manual for more details http://www.austinisd.org/sites/default/files/dept/reach/SLO_Manual_2014_2015interactiveFinal.pdf

[8] See the Jefferson County Public Schools individual educator growth goals guidance for more details https://docs.google.com/a/jeffcoschools.us/file/d/0B-mtlzwdEsvSUzBoZGFsQmNqRHM/edit?pli=1

Conclusions

The source and type of assessment both were related to changes in student performance from pretest to posttest. In general, student scores grew more on common district assessments and on rubric/performance-based assessments. However, the data revealed that student growth was lowest for one specific combination of assessment source and type (teacher-created multiple-choice assessments).

Student growth was comparable for three of the four assessment categories (i.e., common multiple-choice, common rubric/performance-based, and teacher rubric/performance-based). Growth on teacher-created multiple-choice tests was significantly less than the rest. Additionally, fewer students met their individual growth targets and more students demonstrated negative growth on teacher-created multiple-choice assessments than on the other categories.

Future research should continue to understand the relationship between meeting SLO growth targets and earning appraisal points. Specifically, research should investigate differences in appraisal points earned for teachers using the different assessment categories and any impact of differences in appraisal points earned from SLOs on teachers' overall appraisals.

Practical Implications for Teacher Appraisal

Although teacher-created assessments offer the most flexibility to teachers with respect to choice of content and rigor of SLO assessments, common assessments of student growth appear to have a general advantage over teacher-created assessments. If using common assessments, no practical consequences appear to be associated with either multiple-choice or rubric/performance-based assessment types. However, if teachers choose to create their own assessments, they would be wise to create a rubric/performance-based assessment. Teachers may be at a systematic disadvantage regarding the amount of growth their students might achieve when the teacher-created assessment is multiple-choice. Because part of teacher appraisal is based on the percentage of students achieving growth targets (i.e., at least half the distance to a perfect score), it is important for teachers to understand that it may benefit them to use either a common district assessment or rubric/performance-based assessment.

Teachers may be at a systematic disadvantage regarding the amount of growth their students might show when the teacher-created assessment is multiple-choice.

Part of the motivation behind the present study of student growth on SLO assessments was understanding whether the thresholds for meeting growth targets across SLO assessments should be adjusted to prevent systematic bias in measured student growth due to choice of assessment. If bias is present in the percentage of students

meeting growth targets, then potential bias in the points teachers accumulate toward their appraisal should be consider also. Given the comparability of student growth for the common assessments, no adjustment to the thresholds of common assessments is warranted. Although the teacher-created multiple-choice assessments were worse than were other assessments with respect to student growth, percentage of students meeting growth targets, and percentage of students demonstrating negative growth, no adjustment to the thresholds of teacher-created assessments is recommended. Rather, the potential systematic bias in student growth should be acknowledged by teachers who choose to create their own multiple-choice SLO assessments. A multitude of vetted and approved common multiple-choice assessments are available across subjects and levels from the REACH bank of district assessments.

Beyond SLOs: The Broader Conversation

The analyses in this report bring to light some broader issues related to assessing student growth. Among these issues was the use of percentages for assessing student growth, given the various scale ranges across different assessment scales and different pretest scores from which growth was computed. Ultimately, the issues raise the questions of whether growth percentages are truly comparable and when comparability is a concern. Specifically, are the growth percentages computed from different assessment scales comparable, and are the growth percentages computed from the same scales across different ranges comparable? Are there circumstances under which the growth percentages are or are not comparable (e.g., within a classroom and across the district)? These are important contextual conversations to have when considering the use of growth percentages in assessment of students' learning.

Prior SLO Research in AISD

Using teacher-level SLO data, Schmitt (2014) found differences in the percentage of students who met growth targets favoring rubric/performance-based assessments over multiple-choice assessments (i.e., $d > .3$) at both the middle and high school levels.

Using student-level SLO growth data, the current study also found a growth advantage for rubric/performance-based assessments over multiple-choice assessments, but was further able to explore the interaction of assessment type and assessment source. Analysis of the interaction suggested poor performance of teacher-created multiple-choice assessments was responsible for the observed main effect.

Schmitt (2014) also found a general advantage for rubric/performance-based assessments over multiple-choice assessments within the English language arts (ELA) subject (sample availability/size restricted analysis across subjects). As with the study by Schmitt (2014), the current study (see Schmitt & Hutchins, 2015, technical supplement 14.85b) found the same assessment type advantage within ELA, regardless of assessment source. Evidence was mixed regarding any difference between rubric/performance-based and multiple-choice assessments within the common assessment source, only depending on statistical control for pretest performance.

References

Betebenner, D.W. (2008). *Norm- and criterion-referenced student growth* Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publication_PDFs/normative_criterion_growth_DB08.pdf

Schmitt, L. N. T. (2011). *AISD REACH program update, 2010–2011*: *Texas Assessment of Knowledge and Skills growth and student learning objectives* (DRE Publication No. 10.84 RB). Austin, TX: Austin Independent School District.

Schmitt, L. N. T. (2014). *AISD REACH program update: Student learning objective assessments* (DRE Publication No. 13.89 RB). Austin, TX: Austin Independent School District.

Schmitt, L. N. T., & Hutchins, S. D. (2015). *Technical supplement to student learning objectives (SLOs): Analysis of student growth in 2014–2015, by type and source of assessment* (DRE Publication No. 14.85b). Austin, TX: Austin Independent School District.

Schmitt, L. N. T., Lamb, L. M., Cornetto, K. M., & Courtemanche, M. (2014). *AISD REACH program update, 2012–2013: Student learning objectives* (DRE Publication No. 12.83b). Austin, TX: Austin Independent School District.

AUSTIN INDEPENDENT SCHOOL DISTRICT

Author
Lisa Schmitt, Ph.D.
Shaun Hutchins, Ph.D.

Department of Research and Evaluation

1111 West 6th Street, Suite D-350 | Austin, TX 78703-5338
512.414.1724 | fax: 512.414.1707
www.austinisd.org/dre | Twitter: @AISD_DRE

November 2015

Publication 14.85