Learning Sciences
MARZANO
CENTER

Marzano Center
Student Learning Objectives

By Robert J. Marzano,
Lindsey D. Basileo,
and Michael D. Toth

Learning Sciences International
LEARNING AND PERFORMANCE MANAGEMENT

May 2015
West Palm Beach, FL

Our Mission

Learning Sciences Marzano Center, West Palm Beach, Florida, promotes excellence in public education by developing and providing next- generation pedagogical tools, data systems, and training for K-12 educators at the school and district level. Our research department is committed to producing high quality and objective research dedicated to producing evidence-based results, ultimately facilitating sound policy and best practices in education. Under the direction of national researcher and author Dr. Robert Marzano, the Marzano Center identifies, develops, and disseminates cutting-edge resources in public education. With a staff of expert practitioners, consultants, and researchers, our goal is to support all K-12 educators to be highly effective, lifelong learners, and in doing so, to significantly impact student growth and achievement over time.

Michael D. Toth, CEO

Robert J. Marzano, Ph.D.
Executive Director

Visit **MarzanoCenter.com** to learn more.

Table of Contents

6 INTRODUCTION

7 CURRENT STATE

10 THEORY BASE FOR MARZANO CENTER SLOs

10 Tracking Individual Student Growth

11 Advantages of the Growth Model

14 THE ISSUE OF PARALLEL TESTS

14 Unidimensionality

16 Content Articulated at Specific Levels of Complexity

17 Varying Assessment Formats

18 DESIGNING MARZANO CENTER SLOs

19 CONCLUSION

20 REFERENCES

Introduction

Student learning objectives (SLOs) are measurable, long-term academic growth targets that teachers set at the beginning of the school year. Many school districts across the country have been moving toward ways of including multiple measures of teacher performance and student learning into their evaluation systems (Morgan & Lacireno-Paquet, 2013). SLOs are created by teachers and approved by principals. They are long-term, specific learning goals or targets that can be tracked and used to measure teacher impact on student learning (Gill et al., 2013; The Reform Support Network, 2011). In most cases, SLOs are used as performance metrics, setting benchmarks and assessing whether teachers reached the targets set. They are usually based on prior student learning data and aligned to state standards. SLOs can take into account the course subject matter and students with low proficiency (Lacireno-Paquet et al., 2014).

While there are many advantages of using SLOs, especially in non-state-tested subject areas, the current methods of constructing SLOs do not incorporate predictive metrics of student growth. Incorporating these metrics can account for errors found within student growth estimates. Marzano Center Student Learning Objectives (MCSLOs) not only incorporate predictive student growth metrics but also formative assessments and quiz grades to create student learning trajectories for students, ultimately tracking student progress to standards. Moreover, MCSLOs include a reliability estimate that determines the consistency of classroom-level data. This is essential in the implementation of SLOs because districts can set benchmarks not only around student achievement and growth, but also in regard to precision of measurement. The MCSLOs foster teacher autonomy by using formative assessment data while incorporating reliability estimates to ensure consistency in classroom data so that teachers do not inadvertently influence student scores. The aim of this paper will be to detail the MCSLOs.

Current State

As of 2014, there were 30 states that were either using or planning to use SLOs for educator evaluation (Lacireno-Paquet et al., 2014). Of those, 21 states have all teachers implement them or have all teachers implement them within pilot schools; three states apply them only to teachers in specific grades or subject areas in non-tested courses; one state leaves the discretion to the district; and five states do not specify how they are applied to teachers (Lacireno-Paquet et al., 2014). Despite the widespread use of SLOs, there is very little research on their effectiveness at measuring student growth (Gill et al., 2013, 2014; Harris, 2012; Tyler, 2011). Most research documents how to create SLOs and how to implement them with fidelity (Barge, 2013; Center for Assessment, 2013; Indiana Department of Education; Lachin-Hache et al., 2012; Lacireno-Paquet et al., 2014).

SLOs were created as a proxy for measuring student growth in non-tested areas. While it is a requirement for teachers to have student growth data incorporated in their overall teacher evaluation scores, not all teachers have growth data available on the students they teach. SLOs provide a way for teachers to measure how well their students are performing over a given period of time, ultimately satisfying this statutory requirement. SLOs are a status-based model that gives a snapshot of student performance at one point in time and creates targets for student learning for the current year (Marzano and Toth, 2013). Creating SLOs begins by assessing the needs of the students in the class, taking into account baseline data from the previous year, then setting targets that are realistic and challenging for students expected to achieve proficiency by the end of the course (Gill et al., 2014). The process of setting specific goals and measuring achievement provides evidence of each teacher's instructional impact in non-tested areas. It is somewhat a "backwards planning" exercise for teachers as a vision of student success is realized, and then instruction is planned based on that goal (Indiana Department of Education).

Lachlan-Hache and colleagues (2012) describe five types of SLOs that can be created: course-level SLOs that focus on an entire student population for a given course; class-level SLOs that focus on the student population in an individual class; targeted-student SLOs in which separate SLOs are created for subgroups of students who need specific support; targeted-content SLOs in which separate objectives are created for a specific skill or content that students must master; and tiered-targets. SLOs that are often used in conjunction with course-level or class-level SLOs to set differentiated targets based on the range of student abilities in the classroom.

The three best-known examples of SLO implementation are the Austin REACH, Denver ProComp, and Charlotte TIF-LEAP projects (Morgan and Lacireno-Paquet, 2012). These projects have incorporated SLOs into compensation-based models of student performance. The Austin Independent School District (AISD) in Texas is implementing a compensation-based model (REACH) in which teachers develop SLOs to measure student growth. With guidance from the district, teachers are required to develop two SLOs. For one of the SLOs, teachers must achieve the defined target with at least 75 percent of students in a class. For the second, either 75 percent of students in the class or a subgroup of students must reach the defined target. Students are assessed on the objectives at the beginning and end of the year using existing district tests or teacher-developed assessments that are approved by the district. Administrators assess how rigorous the SLOs are by using a predefined rubric. Teachers are required to justify how the SLOs are aligned with student needs, state and national standards, and school improvement plans.

Another example of SLOs is found in Denver. Researchers from Community Training and Assistance Center (CTAC) developed a four-level rubric (4–Excellent, 3–Acceptable, 2–Needs Improvement, 1–Too Little

to Evaluate) to assess the rigorousness and quality of SLOs developed by teachers. The criteria for the rubric levels were determined by a review of teacher planning guides; scope, sequence, and subject standards; and the guidance listed on teachers' objectives forms (CTAC, 2004). After a thorough analysis of the objectives, it was found that students whose teachers developed objectives deemed as "excellent" achieved higher test scores than students whose teachers' objectives were scored lower than "excellent" on the rubric (CTAC, 2004). Moreover, students whose teachers met at least one of their SLOs were more likely to score higher on their standardized assessments than students whose teachers did not meet their SLOs (Goldhaber and Walch, 2012; Schmitt & Ibanez, 2011). While this pilot study found that 70–80% of the teachers met at least one target, a similar study in Tennessee found that roughly two-thirds met all targets (Goldhaber & Walch, 2012; Proctor, Walters, et al., 2011; Tennessee Department of Education, 2012).

The Charlotte TIF-LEAP program was also a compensatory-based model. For this study, researchers found positive, statistically significant associations between how rigorous SLOs were and student achievement, depending on the year of implementation (CTAC, 2013). More specifically, during the second year of implementation the SLO quality rating was positively associated with increased achievement in elementary math and reading and middle school math (CTAC, 2013). Researchers also found that the quality of SLOs increased over time with attainment according to the number of years teachers had participated in the initiative (CTAC, 2013).

There are several advantages in using SLOs. The first major advantage is that teachers in courses or subjects that do not have standardized assessments or in non-tested courses can incorporate a student growth component based on the teachers' students (Gill et al., 2014; Indiana Department of Education; Morgan & Lacireno-Paquet, 2013). Furthermore, SLOs are adaptable and flexible across all subjects and grade levels. They allow for customization and can target students' needs and course goals, connecting to instructional improvement more so than other value-added models. The second major advantage of SLOs is that they can improve collaboration between teachers and principals, pushing teachers to use data to drive instruction, ultimately improving teachers' instructional practice (Gill et al., 2014; Morgan & Lacireno-Paquet, 2013).

Although incorporating SLOs into the teacher evaluation process has clear advantages, it is not without its own set of challenges. The development of SLOs can be very resource-intensive (Gill et al, 2013, 2014). In order for SLOs to be successful, they must be rigorous, and the process of identifying and hitting targets must be investigated, verified, and approved (Morgan & Lacireno-Paquet, 2013). SLOs are also time-intensive (The Reform Support Network, 2011). They require time and support from administrators, training for teachers, and technology to capture objectives, targets, and information regarding whether teachers met those targets (Gill et al., 2013; The Reform Support Network, 2011). Another major challenge in using SLOs is whether they are comparable to value-added metrics and if they are valid and reliable measures of student growth. There have been no studies that reported the reliability of measures of SLO ratings from year to year (Gill et al., 2014). Since individual teachers create SLOs using their own professional judgment, it can be difficult to ensure consistency, and issues with comparability arise (Gill et al., 2013; 2014; Indiana Department of Education; Lachlan-Hache, Cushing, & Biovana, 2012). There is also concern, since teachers have a stake in the results of achieving their SLO targets, there may be incentive to set low expectations or to inflate scores or grades (Gill et al., 2013).

Overall, the limited research on SLOs finds that most teachers achieve at least some of their SLO targets (CTAC, 2013; Goldhaber & Walch, 2012; Proctor et al., 2011; Tennessee Department of Education, 2012; Terry, 2008). Only two studies have correlated SLO ratings with teacher

value-added estimates and have found small positive relationships (Goldhaber & Walch, 2012; Schmitt & Ibanez, 2010). Some researchers note that SLOs can improve student learning (Barge, 2013; Beesley & Apthorp, 2010), but no studies have reported on the reliability of measures of SLO ratings on a year-to-year basis, and more research is needed regarding teachers meeting their SLOs and student net achievement (Gill et al., 2014; Tyler, 2011). While Harris (2012) describes SLOs' potentially attractive qualities of allowing for teaching autonomy in setting individualized objectives and customizing instruction accordingly, he also argues that those same qualities could lend themselves to manipulation and non-comparability. Moreover, he states "there is essentially no evidence about the validity or reliability of SLOs" (Harris, 2012).

The MCSLOs solve most of the issues noted by using a more rigorous measure of student growth incorporating predictive metrics that can be applied to all subjects and courses and by creating reliability estimates for classroom-level data while allowing teacher autonomy to use formative data to track student progress. We begin with the theory base for the Marzano Center approach to SLOs.

Theory Base For Marzano Center SLOs

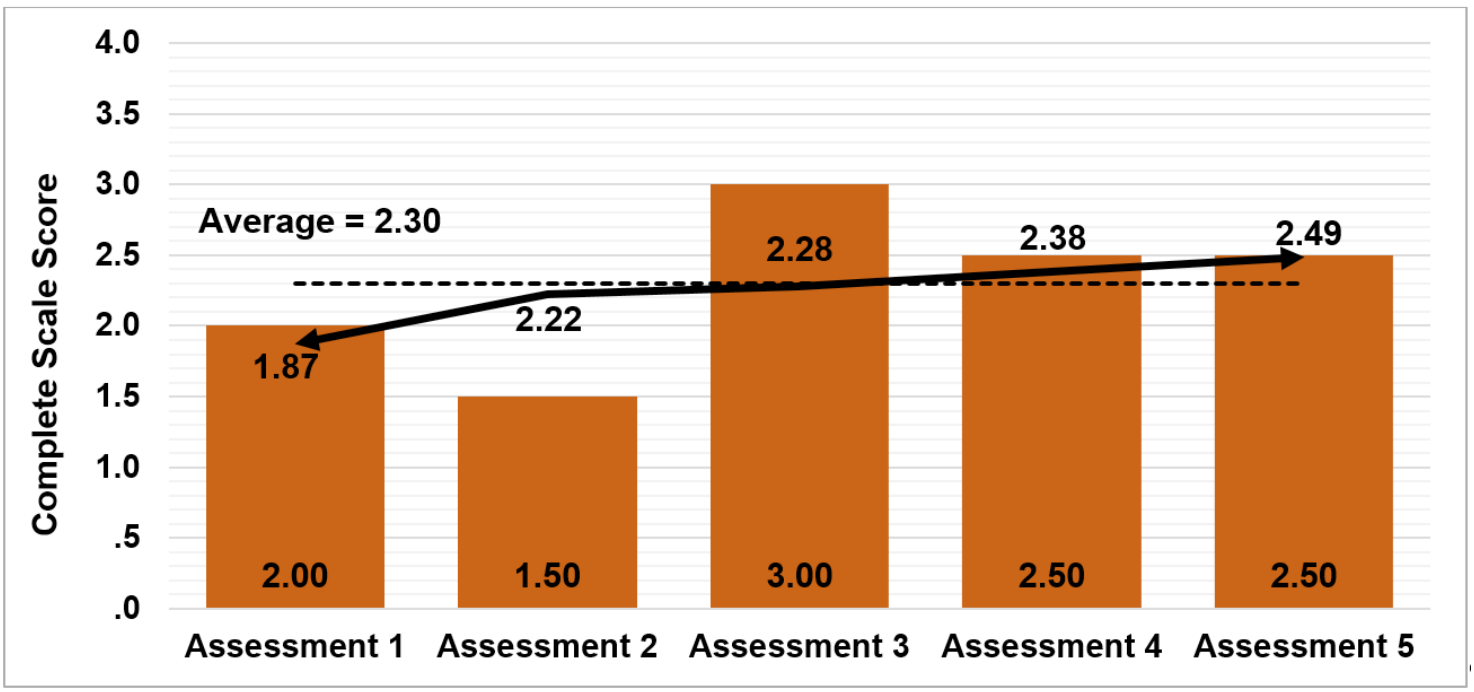
Any effective system of measurement begins with a strong theory base. Indeed, the theory base for the current systems of measurement used in education began at the turn of the last century with the works of Charles Spearman and others (see Traub, 1997 for a discussion). The original theory base (i.e., classical test theory) and its adaptations (e.g., item response theory, generalizability theory) tend to approach the construct of reliability from the perspective of a single, large-scale assessment specifically when it comes to computing a reliability index. The Marzano Center approach focuses on multiple classroom-level assessments collected over time that are focused on specific content. To this extent, MCSLOs are closely related to what Lachlan-Hache and colleagues (2012) refer to as targeted-content SLOs. However, they can also be readily employed as course-level SLOs, class-level SLOs, targeted-student SLOs, and tiered SLOs.

Tracking Individual Student Growth

Perhaps the best way to conceptualize the MCSLO approach is to examine Figure 1, which depicts an individual student's progress across five assessments.

Each student in Figure 1 has received five scores on a specific topic collected over time. The observed scores on these five assessments are depicted by the bar graphs, which represent a scale that ranges from 0 to 4. It is important to note that any scale could be used to create

Figure 1. Individual Student Progress Over Time



an MCSLO. An assumption underlying the depiction in Figure 1 is that all assessments address the same dimension. Stated differently, the assumption is that all assessments are unidimensional.

Unidimensionality is foundational to classical test theory. Lord (1959), in a foundational article on classical measurement theory, noted that a test "is a collection of tasks; the performance on these tasks is taken as an index of his [a student's] standing on some psychological dimension" (p. 473). Thissen and Wainer (2001) note that, "Before the responses to any set of items are combined into a single score that is taken to be, in some sense, representative of the responses to all the items, we must ascertain the extent to which the items "measure the same thing" (p. 10). Unfortunately, many test designers fail to adhere to the unidimensionality criterion (Hattie, 1985; Hattie, Krokowski, Rogers, & Swaminathan, 1996).

Within the MCSLO process, unidimensionality for all assessments in a set is a prerequisite. Additionally, all tests must be designed as equivalent forms. While daunting at first blush, these requirements can be met, at least in part, by adopting a unique perspective on test design at the classroom level. (This perspective is explained in depth in a subsequent section of this paper). Given that each student in a class has been scored multiple times on unidimensional assessments of equivalent form, a graph like that in Figure 1 can be generated for each student and compared for students within and between classes.

A distinguishing feature of the graph in Figure 1 is the reporting of an observed score for each assessment (i.e., the score represented by each bar) and the predicted scores represented by the line cutting through each the bar in the Figure. As described by Willett (1985), the basic measurement model for this endeavor is:

$$X_{ip} = F_p(t_i) + e_{ip}$$

Where the subscript i denotes the occasion of measurement, t_i is the time at which the ith occasion of measurement occurred, and the subscript p indicates the person being measured. The symbol F_p represents the true score status for person p, and the parenthetical inclusion of the time at which the ith measurement occurred indicates that F_p is changing (growing) over time. Thus, the true score for each student at each occasion of measurement is estimated using some hypothesized function (e.g., a linear function) regarding student growth in learning the topic that is the focus of assessment.

This is not a new concept. Willett (1985, 1988) and Rogosa, Brandt, & Zimowsky (1982) have written about it extensively for decades, and Marzano (2006) has discussed the concept in terms of classroom assessment and grading. Angoff (1964) alluded to this concept in the mid-1960s. Specifically, he noted that with successive measurements over time, "we can postulate a single true line and expect random variation . . . to occur about this line" (p. 12). Contrast this measurement model with the traditional model from classical test theory:

$$X_p = F_p + e_p$$

Here F_p represents the true score of person p. The absence of the parenthetical expression (t_i) illustrates that the classical measurement model is restricted to a fixed true score and is mute on the topic of true score change over time.

Advantages of the Growth Model

The growth approach has a number of advantages. One is that each student's true score on each assessment can be readily estimated. To illustrate, in Figure 1, the student's observed score on the first assessment was 2.0, but his predicted true score was 1.85; the student's observed score on the third assessment was 3.0, but his predicted

true scores was 2.28; and so on. Thus, the growth approach allows for estimates of true scores for each student along with the observed scores on each occasion of measurement.

Contrast this with any system that considers individual assessments in isolation, such as one that would use the percentage of students who met or exceeded a specific cut score as the criterion for an SLO. In such cases, the best that can be done regarding estimating the true score for an individual student is to compute a confidence interval around the student's observed score using the formula:

$$\sigma_e = \sigma_x (1 - r_{xx})^{1/2}$$

In this formula, σ_e stands for the standard deviation of the distribution of error scores around a specific observed score; σ_x stands for the standard deviation of observed scores; and r_{xx} stands for the reliability coefficient for the observed scores. To illustrate the application of this formula, assume that a given student receives an observed score of 70 on an assessment that has a reliability of .75 and a standard deviation of 10; the 95% confidence interval would be from 58 to 82.

Another advantage to the growth approach is that it allows for the estimation of reliability when classroom assessments are used. Computing reliability estimates has been the Achilles' heel of classroom assessments since most classroom-based assessments are not designed with an eye toward the psychometric requirements associated with large-scale assessments. Classroom assessments typically do not meet the threshold requirement in terms of quantity of items, item types, and characteristics to be used with traditional formulas for reliability. In effect, traditional formulas for reliability represent an impediment to computing an index estimating the precision of classroom assessments considered as a set.

The problem with using traditional formulas for estimating reliability of classroom assessments is that they are designed to be applied to single assessments. This is a bit curious, since reliability is defined in terms of multiple assessments. As Feldt and Brennan (1993) note, the field of measurement has been preoccupied for years with identifying techniques for computing reliability coefficients using data from a single test even though the concept of reliability is grounded in the concept of multiple administrations of parallel tests:

> **For more than three quarters of a century, measurement theoreticians have been concerned with reliability estimation in the absence of parallel forms. Spearman (1910) and Brown (1910) posed the problem; their solution is incorporated in the well-known formula bearing their names. In the ensuing decades, a voluminous literature has accumulated on this topic. The problem is an intensely practical one. For many tests, only one form is produced, because a second form would rarely be needed. . . . Even when parallel forms exist and the trait or skill is not undergoing rapid change, practical considerations might rule out the administration of more than one form. (p. 110).**

In effect, with the traditional approach, reliability estimates are limited to datasets in which there is one observed score for each test taker. Formulas for computing reliability on single assessments are designed for large-scale assessments with many items.

A growth approach changes the perspective to datasets with multiple scores for each test taker collected over time. It is not the length of a single test that creates precision; it is the number of data points for each subject collected over time that creates precision. This fits classroom assessment quite well. True scores for each test-taker on each test can be estimated assuming some specific growth function over time. The example in Figure 1 assumes a linear function to estimate the true scores; other functions (i.e., power functions, exponential function, and so on) can all be used. The basic equation for a reliability estimate on multiple assessments over time is:

$$p(\hat{B}) = \frac{\sigma_B^2}{\sigma_B^2 + \frac{\sigma_e^2}{SST}}$$

Here σ_B^2 represents the variance of growth rates; σ_e^2 represents the variance due to measurement error; SST represents the sum of squares total. To illustrate the application of this formula, consider the matrix in Table 1, which contains the scores of 10 students over five assessments.

The reliability of the scores for these 10 students is .881. This reliability is not only quite high, relatively speaking, but more importantly it is calculated using classroom assessments that, if examined in isolation, might have relatively low reliabilities calculated using formulas that are designed to be applied to individual assessments.

Table 1. Students Observed and Predicted Scores With Beta Weights

Student	Score	Assessment 1	Assessment 2	Assessment 3	Assessment 4	Assessment 5	Beta
Student 1	Observed	2.50	2.00	2.00	2.50	3.50	.693
	Predicted	1.96	2.31	2.43	2.70	3.09	
Student 2	Observed	1.50	2.50	2.50	3.00	4.00	.990
	Predicted	1.56	2.30	2.55	3.13	3.95	
Student 3	Observed	2.50	3.50	3.00	3.50	3.00	.389
	Predicted	2.89	3.03	3.07	3.18	3.33	
Student 4	Observed	1.50	2.00	2.50	3.50	2.50	.662
	Predicted	1.78	2.18	2.32	2.63	3.09	
Student 5	Observed	1.50	2.00	2.00	1.50	2.50	.653
	Predicted	1.55	1.78	1.85	2.03	2.28	
Student 6	Observed	2.00	3.00	3.00	2.50	3.50	.757
	Predicted	2.25	2.61	2.73	3.01	3.40	
Student 7	Observed	3.00	3.00	2.50	3.00	3.50	.552
	Predicted	2.75	2.91	2.97	3.09	3.27	
Student 8	Observed	1.00	1.50	2.50	3.00	2.50	.785
	Predicted	1.28	1.82	1.99	2.41	3.00	
Student 9	Observed	3.50	2.50	2.50	3.00	4.00	.408
	Predicted	2.76	2.98	3.06	3.23	3.47	
Student 10	Observed	2.50	2.50	3.00	3.50	3.00	.653
	Predicted	2.55	2.78	2.85	3.03	3.28	
Average Teacher Growth							.654

The Issue of Parallel Tests

Multiple administrations of parallel assessments that are parallel are central to the general concept of reliability and critical to the calculation of the reliability of assessments administered over time. Unfortunately, the theory and practice of parallel assessments are complex issues. As recently as 1966, Horst noted that "the assumptions underlying the definition and construction of parallel test forms have not been adequately set forth" (p. 295).

This observation notwithstanding, Traub (1997) noted that discussions of parallel forms have a long history, especially as they relate to the concept of reliability. Specifically, between 1910 and 1925, the index of reliability was commonly conceived as the correlation between repeated measures of the same or identical tests. The reasoning underlying such discussions usually operationally defined reliability as the correlation between parallel tests and then operationally defined parallel tests. For example, in 1940 Gulliksen noted that "we shall define reliability as the correlation of parallel forms of a test" (p. 13). He then noted:

> **Instead of defining parallel tests in terms of true scores and error (as we did in the chapter immediately preceding) then deriving the observed score characteristics of parallel tests, we shall define parallel tests in terms of observed score characteristics. (p. 29).**

Gulliksen then listed a number of characteristics of observed scores of parallel tests, such as equal standards deviations, equal pairwise correlations, and so on.

This last point of Gulliksen's is very germane to the present discussion—namely, parallel tests are usually defined in terms of the psychometric properties of the test items and the observed scores they generate. From this perspective, it would be difficult, if not impossible, for an individual teacher or set of teachers to construct parallel tests since they do not have the resources (i.e., time, energy, access to statistical packages) and in some cases the technical expertise to do so. For the MCSLO approach, we offer a different perspective on parallel tests that can be designed by groups of practitioners. This approach is grounded in three principles: unidimensionality, content articulated at specific levels of complexity, and varying assessment formats.

Unidimensionality

The importance of unidimensionality in measuring student growth was discussed previously. In short, it has been a defining characteristic of parallel tests. As Hambleton (1993) noted, "The notion of an underlying latent ability, attribute, factor, or dimension is a recurring one in the psychometric literature" (p. 149). Interestingly, many test designers fail to adhere to the unidimensionality criterion (Hattie, 1985; Hattie, Krakowski, Rogers, & Swaminathan, 1996).

Within the MCSLO process, unidimensionality is achieved by the articulation of a learning progression in the form of a proficiency scale. The concept of a learning progression became popular about the same time as discussions about formative assessment. Heritage (2008) explained the link between learning progressions and formative assessments as follows:

> **The purpose of formative assessment is to provide feedback to teachers and students during the course of learning about the gap between students' current and desired performance so that action can be taken to close the gap. To do this effectively, teachers need to**

have in mind a continuum of how learning develops in any particular knowledge domain so that they are able to locate students' current learning status and decide on pedagogical action to move students' learning forward. Learning progressions that clearly articulate a progression of learning in a domain can provide the big picture of what is to be learned, support planning, and act as a touchstone for formative assessment. (p. 2).

To illustrate, consider the progression for the concept of buoyancy designed by Herman and Choi (2008):

- **Student knows that floating depends on having less density than the medium.**
- **Student knows that floating depends on having a small density.**
- **Student knows that floating depends on having a small mass and a large volume.**
- **Student knows that floating depends on having a small mass or that floating depends on having a large volume.**
- **Student knows that floating depends on having a small size, heft, or amount, or that it depends on being made out of a particular material.**
- **Student thinks that floating depends on being flat, hollow, filled with air, or having holes.**

This progression of knowledge can be organized into a scale like the one depicted in Table 2.

Assessments can be designed to address the various levels of the scale and that are then administered and scored by teachers. In effect, the scale establishes a blueprint for assessment design that helps ensure unidimensionality for all assessments that are based on the explicit content described at the various levels. (For a discussion on how assessments are constructed using such scales, see Marzano, 2010.)

The learning progression above was designed empirically. When such empirically based progressions are not available, approximations to them can be designed by teachers working in collaborative teams. To this end, Marzano (2010) has recommended the use of proficiency scales. Proficiency scales require teachers to identify at least three levels of explicit content. Table 3 depicts a teacher-designed proficiency scale for eighth-grade content about Napoleon.

Table 2: Buoyancy Progression Organized as a Scale

Score 4.0	Student knows that floating depends on having less density than the medium
Score 3.5	Student knows that floating depends on having a small density
Score 3.0	Student knows that floating depends on having a small mass and a large volume
Score 2.5	Student knows that floating depends on having a small mass, or student knows that floating depends on having a large volume
Score 2.0	Student thinks that floating depends on having a small size, heft, or amount, or that it depends on being made out of a particular material
Score 1.5	Student thinks that floating depends on being flat, hollow, filled with air, or having holes
Score 1.0	With help, partial success at score 2.0 content and score 3.0 content
Score 0.5	With help, partial success at score 2.0 content, but not at score 3.0 content
Score 0.0	Even with help, no success

The articulation of a proficiency scale helps ensure unidimensionality since it makes it easy to ascertain that all content at each level relates to the same ability, attribute, factor, or dimension. In essence, a proficiency scale turns an otherwise latent ability, attribute, factor, or dimension into explicit elements.

Table 3: Proficiency Scale for Napoleon

Score 4.0	The student: • Compares and contrasts Napoleon and other military and political leaders. *No major errors or omissions regarding the score 4.0 content*
Score 3.5	In addition to score 3.0 performance, partial success at score 4.0 content
Score 3.0	The student: • Makes a flowchart depicting the rise and fall of Napoleon (e.g., creates an illustrated flowchart that includes Napoleon's 1799 coup, his major military achievements, and his final invasion of Russia) *No major errors or omissions regarding the score 3.0 content*
Score 2.5	No major errors or omissions regarding score 2.0 content, and partial success at score 3.0 content
Score 2.0	The student: • Recalls accurate information about the rise and fall of Napoleon, such as: • He was not French by birth and never mastered the language • His first position of significant military command was with France's Army of Italy • He was imprisoned and then exiled to the island of St. Helena in 1815 *No major errors or omissions regarding the score 2.0 content*
Score 1.5	Partial success at score 2.0 content, and major errors or omissions regarding 3.0 content
Score 1.0	With help, partial success at score 2.0 content and score 3.0 content
Score 0.5	With help, partial success at score 2.0 content, but not at score 3.0 content
Score 0.0	Even with help, no success

Content Articulated at Specific Levels of Complexity

Virtually all discussion of parallel tests addresses the requirement that the distribution of item difficulties should be similar if not identical. This is difficult for classroom teachers to accomplish. However, the construction of a proficiency scale helps teachers design assessments that include content at each level of the scale. To illustrate, consider the proficiency scale above. All assessments designed with this scale as a reference point would include content from the three explicit content levels of the scale (i.e., scores 2.0, 3.0, and 4.0). Under the assumption that the levels of a scale represent different levels of content difficulty, the assessments designed using the scale could reasonably approximate the requirement of similar item difficulty distributions across assessments.

Varying Assessment Formats

The third principle underlying the MCSLO process is that assessments would be allowed to manifest in different formats. Before addressing this issue, it is useful to provide some working definitions for understanding constructs of the MCSLO process. These are presented in Table 4.

When a proficiency scale has been created, terms like assessment, measurement, scale, and score take on specific meanings. An assessment becomes any systematic method a teacher uses to draw inferences about a student's position on the proficiency scale at a particular moment in time. The score a student is assigned on a particular assessment at a particular time is always a point on the proficiency scale. The proficiency scale itself serves as the scale with which all assessments are interpreted, since it is a system of numbers and their units by which a value is reported on some dimension. Finally, the process of translating students' results on various assessments into points on the proficiency scale is by definition the process of measurement. In effect, teachers could use different types of assessment formats but still reference the same scale. For example, a score of 2.0 on any assessment referenced to the scale always means the same thing in terms of students' levels of expertise.

Table 4: Definitions

Assessment	Any systematic method of obtaining information used to draw inferences about characteristics of people, objects, or programs; a systematic process to measure or evaluate the characteristics or performance of individuals, programs, or other entities for purposes of drawing inferences; sometimes used synonymously with test (AERA, 2014, p. 216).
Measurement	The assignment of numerals to objects or events according to rules (Stevens, 1946, p. 677).
Scale	The system of numbers and their units by which a value is reported on some dimension of measurement. In testing, the set of items or subsets used to measure a specific characteristic (e.g., a test of verbal ability or a scale of extroversion-introversion) (AERA, 2014, p. 223).
Score	Any specific number resulting from the assessment of an individual, such as a raw score, a scale score, an estimate of a latent variable, a production count, an absence record, a course grade, or a rating (AERA, 2014, p. 223).

Designing Marzano Center SLOs

The process of designing MCSLOs begins with the six steps described in Table 5.

Table 5: Six-Step Process Underlying the Construction of SLOs

Step 1	Create a proficiency scale for a unit of instruction that will be taught by a group of teachers. The unit might be as short as a few weeks or as long as a grading period.
Step 2	Create a common pretest (one that will be administered by all teachers) and a common posttest using the proficiency scale.
Step 3	Administer the common pretest at the same point in time for all teachers. If possible, multiple teachers should score common pretests.
Step 4	Allow teachers to create their own interim assessment under the restriction that every assessment provides data that allow for the assignment of scores along all points of the proficiency scale.
Step 5	At the end of the unit, have teachers administer the common posttest at the same point in time. If possible, multiple teachers should score posttests.
Step 6	For each teacher, compute: 1) each student's prediction line (i.e., slope), 2) the average slope for the class, and 3) the reliability of the scores for the students as a set.

The data generated from these six steps are comparable from teacher to teacher within the set using the proficiency scale as the basis of assessment. Average slopes can be compared, as can the reliability of each teacher's measurement process. Additionally, the percentage of students below or above a specific slope can be compared. For example, for each teacher the percentage of students below a slope (expressed as a beta weight) of .10 could be computed; similarly, the percentage of students above a slope of .30 (expressed as a beta weight) could be computed. Teachers could set their own goals relative to expected average slopes for their class, expected reliability for the measurement process, and expected percentages of students below or above specific beta weights.

Conclusion

The approach described above is a new and unique way of designing and implementing SLOs. It uses teacher-designed assessments in such a way that average student growth rates can be computed and aggregated across teachers. Additionally, the precision of teacher assessments can be estimated and used to interpret the utility of individual student scores as well as aggregated scores. The MCSLO process outlined in this paper improves upon the traditional SLO process. MCSLOs incorporate predictive metrics of student growth that reduce errors in estimates. MCSLOs fosters teacher autonomy by using formative assessment data while incorporating reliability estimates to ensure consistency in classroom data. Benchmarks can be set in a number of different ways, not only centered on student growth but also around reliability. If implemented properly, these methods should help teachers track progress of students to standards and ultimately increase student achievement.

References

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1964). Technical problems of obtaining equivalent scores on tests. *Journal of Educational Measurement.* 1, 11–13.

Barge, J. (2013). *Student learning objectives as measures of educator effectiveness: A guide for principals.* Georgia Department of Education. Retrieved from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/SLOs%20Guide%20for%20Principals%20_1-2-2013.pdf

Beesley, A., & Apthorp, H. (2010). *Classroom instruction that works.* Second edition: Research report. Denver, CO: Mid-continent Research for Education and Learning.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology,* 3, 296–322.

Community Training and Assistance Center. (2004). *Catalyst for Change: Pay for Performance in Denver Final Report.* Boston, MA: Author.

Community Training and Assistance Center. (2013). *It's More than Money: Teacher Incentive Fund— Leadership for Educators' Advanced Performance.* Charlotte-Mecklenburg Schools. Boston, MA:Author.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed.). (pp. 105–146). London: Macmillan.

Gill, B., Bruch, J., & Booker, K. (2013). *Using alternative student growth measures for evaluating teacher performance: What the literature says.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2013002.pdf

Gill, B., English, B., Furgeson, J., & McCullough, M. (2014). *Alternative student growth measures for teacher evaluation: Profiles of early-adopting districts.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://files.eric.ed.gov/fulltext/ED544797.pdf

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review,* 31(6), 1067–1083.Retrieved from http://eric.ed.gov/?id=EJ989067

Gulliksen, H. (1940). *Theory of mental tests.* New York: John Wiley & Sons.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd Edition, pp. 147–200). London: Macmillan.

Hattie, J. (1985). Methodology review: Assessing the unidimensionality of tests and items. *Multivariate Behavioral Research,* 19, 49-78.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement,* 20(1), 1–14.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment.* Washington, DC: Council of Chief State School Officers.

Herman, J. L., & Choi, K. (2008 August). *Formative assessment and the improvement of middle school science learning: The role of teacher advocacy.* (CRESST Report 740). Los Angeles: University of California Graduate School of Education and Information Studies, National Center for Research and Evaluation, Standards, and Student Testing (CRESST).

Horst, P. (1966). *Psychological measurement and prediction.* Belmont, CA: Wadsworth Publishing Company.

Indiana Department of Education. *RISE Evaluation and Development System. Student Learning Objective Handbook.* Retrieved from http://www.riseindiana.org/sites/default/files/files/Student%20Learning/Student%20Learning%200bjecObjec%20Handbook%201%200%20FINAL.pdf

Lachlan-Hache, L., Cushing, E., & Biovana, L. (2012). *Student learning objectives-benefits, challenges, and solutions.* Washington, D.C.: American Institutes for Research. Retrieved from http://westcompcenter.org/wp-content/uploads/2013/04/SLO-Benefits-and-Challenges-American-Institutes-for-Research.pdf

Lachlan-Hache, L, Cushing, E., & Bivona, L. (2012). *Student learning objectives as measures of educator effectiveness the basics.* Washington, DC: American Institutes for Research. Retrieved from http://www.ok.gov/sde/sites/ok.gov.sde/files/documents/files/Exploring%20Student%20Learning%200Objective.pdf

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 North-east & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2014013.pdf

Lord, F. M. (1959, June). Problem in mental tests theory arising from errors of measurement. *Journal of the American Statistical Association,* 54(286), 472–479.

Marzano, R. J. (2006). *Classroom assessment and grading that work.* Alexandria, VA: ASCD.

Marzano, R. J. (2009). *Designing and teaching learning goals and objectives.* Bloomington, IN: Marzano Research Laboratory.

Marzano, R. J. (2010). *Formative assessment and standards-based grading.* Bloomington, IN: Marzano Research Laboratory.

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.

Morgan, C., & Lacireno-Paquet, N. (2013). *Overview of student learning objectives (SLO): Review of the literature.* Waltham, MA: Regional Educational Laboratory at EDC.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1993). Scaling, norming, and equating. InR. L. Linn (Ed.), *Educational Measurement* (3rd Ed.,pp. 221–262. Phoenix, AZ: The Oryx Press.

Proctor, D., Walters, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011). *Making a difference in education reform: ProComp external evaluation report, 2006–2010.* Denver: The Evaluation Center.

Rogosa, D. R., Brandt, D., & Zimowsky, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin,* 90, 726–748.

Schmitt, L., & Ibanez, N. (2011). *2009–2010 Texas assessment of knowledge and skills results and student learning objectives.* Austin, TX: Austin Independent School District.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*. 3, 271–295.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science,* 103, 677–680.

Tennessee Department of Education. (2012). *Teacher evaluation in Tennessee: A report on year 1 implementation.* Retrieved from http://eric.ed.gov/?id=ED533726

Terry, B. D. (2008). *Paying for results: Examining incentive pay in Texas schools.* Austin, TX:Texas Public Policy Foundation. Retrieved from http://www.broadeducation.org/asset/1128-paying%20for%20results.pdf

The Reform Support Network. (2012). *A quality control toolkit for student learning objectives.* Washington, DC: Author. Retrieved from https://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/slo-toolkit.pdf

The Reform Support Network (2011). *Targeting growth: Using student learning objectives as a measure of educator effectiveness.* Washington, DC: Author. Retrieved from http://msde.state.md.us/tpe/TargetingGrowth_Using_SLO_MEE.pdf
Thissen, D., & Wainer, H. (2001). An overview of test scoring. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 1- 19). Mahwah, NJ: Lawrence Erlbaum Associates.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice,* 16(4), 8-14.

Willett, J. B. (1985). *Investigating systematic individual difference in academic growth.* (Unpublished doctoral dissertation). Stanford University, Palo Alto, CA.

Willet, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education,* 15, 345–422. Washington, DC: American Educational Research Association.

Learning Sciences
MARZANO
CENTER

Learning Sciences International
LEARNING AND PERFORMANCE MANAGEMENT

1.877.411.7114
MarzanoCenter.com

West Palm Beach, FL