Considerations for Analyzing Educators' Contributions to Student Learning in Non-tested Subjects and Grades with a Focus on Student Learning Objectives

Scott Marion, Charlie DePascale, Chris Domaleski, Brian Gong, and Elena Diaz-Bilello[1]

Center for Assessment

May 25, 2012

Introduction

Many state and district education leaders are rapidly reforming their educator evaluation systems with the results from student performance measures playing a significant role in the overall evaluation results. When conceiving these reforms, it appears that policy makers have not considered the huge challenge of including all teachers in these systems even though only one-quarter or so of the teaching force is in grades and subjects with at least two years of state test data. As discussed previously (Buckley & Marion; Marion & Buckley, 2011), many of the RTTT applications contained promises (or hopes) that states would use other forms of data in order to incorporate student performance results into the evaluations of teachers in "non-tested subjects and grades" (NTSG) or those not assessed by state standardized tests. Student learning objectives (SLO) have gained popularity as a means for attributing student performance results to educators in new forms of teacher evaluation systems for all teachers, but especially for those in NTSG.

SLO are content- and grade/course-specific measurable learning objectives that can be used to document student learning over a defined period of time. To boil SLO down, they provide a means for educators to establish learning goals for individual or groups of students, monitor students' progress toward these goals, and then evaluate the degree to students achieve these goals. The active involvement of the teacher throughout the process is a key advantage of the SLO approach over traditional test-centered approaches to accountability. It is designed to reflect and incentivize good teaching practices such as setting clear learning targets, differentiating instruction for students, monitoring students' progress toward these targets, and evaluating the extent to which students have met the targets.

By their very nature, SLO are focused on the change in student performance over a specified period of time. In fact, good SLO should incorporate some type of stretch goals for students, which obviously means that educators must have some sense of where students start and what would constitute an appropriate stretch. This has led many to rush to the conclusion, supported by USED guidance (Goe & Holdheide, 2011; Secretary's Priorities for Discretionary Grant

[1] The authors acknowledge the thoughtful and useful feedback provided by Derek Briggs.

Programs, 2010), that SLO must be evaluated using "growth-based" measures such as value-added modeling (VAM) or simple gain scores to estimate growth scores for individual students or classes. In practice, however, the focus of evaluation of most SLO will be "status-based"; evaluating the degree to which students reach specific targets at the end of the instructional period. The role of student growth is embedded within the process of establishing performance targets for groups of students depending on some rough sense of where they start, rather than in the technical measurement of change in student performance. Understanding the distinction between embedding growth within the development of the SLO and attempting to measure student growth through the SLO process is critical to the design of appropriate measures to evaluate SLO and ultimately to the feasibility and success of the SLO approach.

However, even understanding this distinction may not be sufficient to overcome the current euphoria over growth as the panacea for student, school, and teacher accountability. Some may simply be tempted to measure growth through the SLO process to be consistent with measures being employed with the large-scale state assessment. This brief is designed to strongly recommend that for educators in most subjects and grades, growth-based measures should be avoided as a means of evaluating SLO. This is particularly true for subjects and grades in which there is not a standardized, large-scale district or state assessment. The reasons for this recommendation are as follows:

- **Technical capacity**. Measuring student growth is more complex than it appears. Most districts lack the capacity to use anything more than the simplest means of calculating growth, such as gain scores, and such approaches may be very unfair in accountability systems because schools or teachers may be held to very different standards.
- **Low quality assessments**. There is a lack of assessments of even reasonable quality in most NTSG. This will lead to the temptation to simply use commercial products even if they are not appropriate for the specific objective (or course).
- **Weak approaches to measure growth in the SLO context**. Most attempts to measure student growth will inevitably come to rely on either sophisticated growth measures (e.g., VAM) or the apparent simplicity of a pretest-posttest design.
    - As stated above, most districts lack the capacity and size to implement sophisticated growth measures. Even in cases where those sophisticated measures can be employed, however, the reliability of those approaches at the teacher level appears to be quite low based on initial research.
    - A pretest-posttest design (within the same course) has been proposed in many cases and there are many well-documented challenges to such designs, including corruptibility, lack of clear conceptualization about the nature of the pretest, weak analytic approaches (especially gain scores), and treated non-equated tests as if they shared the same score scale.

Prior to elaborating on each of these recommendations grouped into two major topics—design/technical issues and assessment quality—it is important to clarify a few points first. The recommendations in this paper apply to most approaches for measuring "growth" in an SLO context, whether referring to a simple gain score approach where the difference between two scores is used as the accountability score or more sophisticated statistical approaches. In contrast, "status" approaches evaluate the extent to which students achieved a specific score/level on a particular assessment. In an SLO framework, the evaluation of achievement may be conditioned (i.e., adjusted) on some rough starting points for the students. This is a recommended approach for analyzing SLOs that is discussed in the last section of this paper. We argue that this might be as close to "growth" as we can get when using SLOs. To be fair, many of these concerns are not eliminated—particularly assessment quality—when using status-based approaches, but relying on growth measures to document student performance compounds these concerns.

Design and Technical Issues

Analytic capacity and gain scores

The most important reason for arguing against a pretest-posttest design is that in spite of the intuitively appealing simplicity of gain scores, it is very difficult to construct a valid pretest-posttest design. Implementing a pretest-posttest design requires considerable technical capacity of school and district personnel to analyze the data in technically appropriate ways. This is not meant to disparage district and school leaders, but simply acknowledges that very few districts have people with enough technical training to use such techniques as VAM or more simple alternatives such as analysis of covariance (ANCOVA) or multiple regression. Therefore, most pretest and posttest data will be analyzed using simple "gain scores," where the pretest score is subtracted from a posttest score. This may happen in spring-to-spring designs for NTSG, but we suspect it will be more prevalent for fall-to-spring designs. Some readers might ask, "what's the problem with gain scores?" There are two fatal flaws with using simple gain scores in educator evaluation systems. The first is that judgments of "gain" are usually based on non-equated test scores and the second is that gain scores do not take into account differences in where students start, which leads to a potentially very unfair accountability system.

Al Beaton once famously said, "if you want to measure change, don't change the measure" when referring to monitoring score changes on NAEP. Likewise, in order to make valid psychometric judgments about longitudinal student growth or improvement across cohorts of students, the scores must be placed on the same scale such that a score of X_1 on one test is "interchangeable" with a score of X_2 on a second test. It is impractical to not "change the measure" with accountability tests, so fairly complicated linking (equating) designs are used to place the scores of one test on the scale of the other test. Such linking is rarely done when using non-commercial tests in NTSG. The default method is often then a very simple gain score approach such as subtracting the fall test from the spring test score (usually a percent correct or rubric score

metric) on non-linked tests. In these cases, growth inferences are almost always invalid. In fact, in a recent study of such practices in a large urban district, Diaz-Bilello (2011) found that "growth" was usually underestimated because the spring test was actually more difficult than the fall test, but these differences in difficulty were not taken into account because an appropriate linking approach was not used. Granted, these psychometric concerns are only true if one wants to know "how much change" in an interval or ratio sense. Most curricular-based indicators of progress are more categorical in nature because users inherently recognize that it makes more sense not to try to have a quantitative vertical scale between the starting point and the measurement point. Ideally, SLO would be designed more from a curriculum-based approach, but unfortunately most rely on attempts to mimic technical approaches to measure growth.

Even if appropriate linking designs are used, simple gain scores are generally not fair because students tend to grow at very different rates regardless of the quality of teaching. Much of the recent research using VAM and SGP as well as a long history of norm-referenced testing has documented the varying rates at which student grow and progress through school and this often contingent upon factors outside of school. A more popular example of this phenomenon involves Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP), which is one of the most, if not the most, popular interim assessment program in the country (Marion, 2009). NWEA provides a "growth expectation" for each student tested based on (i.e., conditioned on) the fall test score and the grade level of the student. The expectation is the median score of students with the same fall score and grade level using historical data from NWEA's extensive database[2]. So what does this mean for teacher evaluation? If a simple average gain score is used as the metric, then unless students are randomly assigned to teachers, the system will be unfair to teachers assigned a group of students likely to grow less than other students. While VAM and related approaches attempt to statistically account for non-experimental conditions, it has been well-documented that even such sophisticated approaches fall short (e.g., Briggs, in press; Rothstein, 2009), but the inferences about teacher contributions to student achievement are even more tenuous with simple gain score approaches.

Problems with pretest-posttest designs

Depending on the type of analytic model used, the conceptualization of a prior score or prior information will change. Much of the language in NTSG currently (e.g., Goe. & Holdheide, 2011), even in the more idiosyncratic subjects and grades (e.g., high school environmental science, French 4), promotes the use of a pretest given early in the course to serve as the "baseline" against which the end of year assessment is evaluated. Evaluations of Title I programs in the early years of ESEA used a pretest-posttest design, but moved away from such approaches in light of shortcomings of this design. Using a pretest in a course makes very good

[2] For more details, see: http://www.nwea.org/support/article/535/growth-norms

sense in an instructional context, but is fraught with problems when used for teacher accountability.

One of the reasons that Title I evaluations moved away from the pretest-posttest design is over concern of what appeared to be inflated fall-to-spring gains. This is more of an issue with younger compared to older students, but the summer "fall back" problem has been well-documented (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996) and contributes to what appear to be larger gains on fall-to-spring tests rather than spring-to-spring. While a case could be made that the fall-to-spring change is a more appropriate for teacher evaluations than spring-to-spring changes, but since most school accountability analyses and large-scale VAM or SGP analyses for tested grades are based on spring-to-spring data, using different frameworks for NTSG could introduce unintended incoherence into the system.

While it might be the pink elephant in the room, it is important to point out the increased likelihood, or at least temptation, of corruption when the teacher is administering both the pretest and posttest. This is not to say that teachers will overtly cheat (although some surely will), but it is easy to imagine scenarios where teachers may tell their students not to worry too much about the pretest since it "doesn't count," while knowing that it certainly counts in their evaluations. This is probably not the most important limitation of the pretest-posttest design, but it is certainly requires acknowledgment. Again, the potential for corruptibility is not limited to pre-post designs, but such designs present additional likelihood of corruption beyond approaches focused on end-of-year assessments. To be fair, SLO offer as many or more opportunities for corruption as any other method, but because SLO should be fully integrated into classroom practice, they perhaps provide greater opportunities than more external approaches for incentivizing positive behaviors. There is also the possibility for both SLO and other NTSG methods that credible external methods (e.g., scores from a prior year in a related course) may be used as either a formal or less formal conditional variable to adjust expectations based on prior achievement.

Unreliability of growth measures

Every measurement, whether physical, psychological, or educational, contains some degree of "error" associated with the scores that are reported. In fact, the field of educational measurement was created to help understand and quantify the differences between observed and "true" scores. In other words, we know a lot about score reliability and standard errors of measurement (SEM). For example, we are sure that all things being equal, longer tests are more reliable than shorter tests. Related to our interest in measuring growth, we also know that "error" associated with the difference in two test scores (e.g., growth) is greater than the error on either test. In fact, in many cases, this error is often larger than the educational change we hope to see. In other words, we would have a hard time distinguishing between real change (either up or down) and the error of the difference score.

In the measurement of growth at the classroom level, we are also concerned about another source of error. There is some debate in the field whether this other source of uncertainty should be called "sampling error" or not, but we do know that there is fluctuation in test and growth scores that behaves similar to sampling error. Sampling error refers to score variability due to differences in the students across classes. Most educators are well aware of the "good class/bad class effect" that describes the changes in class performance based on nothing more than the particular group of student in the class that year. Therefore, one needs to consider this "sampling error" in addition to the measurement error described above.

Research from the Measures of Effective Teaching (MET) project and other studies indicate that using highly reliable tests and sophisticated Value-Added Model (VAM) approaches still results in consistency of VAM results across years at the teacher level of between 0.2-0.4. This is on a scale where 1.0 is perfectly reliable or consistent and 0 means that the two tests are completely unrelated. To put it in context, the consistency of status scores across years tends to be about 0.8 or more for all schools in Colorado, but significantly lower when school sizes approach the sizes of even large classrooms (e.g., Linn & Haug, 2002). In other words, even in the best cases where both the measures and the growth calculations are generally regarded to be technically strong, growth outcomes at the class level are simply not stable. One can only imagine how much more problematic it is to attempt to produce a credible growth results at the class level when conditions are far from ideal.

Assessment Quality

Many have raised concerns about the quality of assessments available to teachers and others for evaluating student learning (e.g., Shepard, 2000; Perie, Marion, & Gong, 2009) and for use in educator evaluation systems (e.g., Herman, Heritage, & Goldschmidt, 2011). There is no question that the lack of high quality assessments in the NTSG has led districts like Hillsborough, FL and Harrison, CO, and states such as Delaware to embark on a massive assessment development efforts. While the quality of these newly created assessments may be better than what was in place previously, there is still a long way to go to have high quality assessments in even a significant percentage of NTSG. Given this challenge of finding at least one good assessment in each of the multitude of NTSG, it just does not seem to make sense to think that we will be able to now find two (pretest and posttest). We note that entities like those mentioned as examples here appear to be moving to a situation where all grades are "tested grades" and a discussion of these approaches may not fit the current discussion of the use of SLO in NTSG. For now, we will leave for later the question of whether it is worth it to pursue such an ambitious external-to-the-classroom initiative, even if it could be done well.

Use of commercial products

States and districts often consider incorporating commercial norm-referenced and/or interim assessments in their educator evaluation systems. There are many legitimate reasons for doing so. In spite of significant concerns about the quality of such assessments (e.g., Perie, et al., 2009), most are scaled and equated using defensible psychometric approaches so that scores across multiple test forms can be compared validly. Further, these assessments are common across the system, which fulfills a policy desire of holding educators to comparable expectations. Therefore, there is the temptation to use such assessments in educator evaluation systems. There are numerous concerns with such a relatively simple rush to use an existing product. An important consideration in any accountability system is that people should only be held accountable for those things that they can control. Most current commercial products are often weakly aligned to what educators are being asked to teach. This difference between the curriculum expectations and the assessments will turn into a chasm as schools more fully implement the Common Core State Standards (CCSS) unless these products are revised considerably. Holding teachers accountable for the results on commercial assessments could hinder the implementation of the CCSS because educators will naturally gear their instruction toward that which they are being held accountable.

SLOs are based on having educators establish goals about important and specific learning outcomes, what Wiggins & McTighe (1998) have termed "enduring understandings." Most norm-referenced and interim assessments tend to cover a broad range of learning outcomes at a fairly superficial level. Therefore, it is unlikely that they will be able to provide an appropriate evaluation of such enduring understandings. Yes, these commercial assessments tend to be fairly reliable measures, but rushing to use such assessments in an educator evaluation system is the proverbial hammer in search of a nail. This is not to say that such commercial assessments should never be included in educator evaluation systems, but only if they meet the critical alignment requirements and users can document that they are appropriate measures of the learning objectives.

In short, if the central purpose of educator evaluation is to promote good instruction and, ultimately, improve student achievement, it is critical that any measures used to draw inferences about teacher efficacy represent what teachers should be teaching and students should be learning. To do otherwise risks incentivizing instructional practices that are not valued – which is to pursue the right answer to the wrong question.

Recommendations

What are some solution-oriented recommendations given all of these cautions? First, while there are many challenges with the implementation of SLO, they offer the most promise for documenting the student performance from NTSG while offering great potential for improving education. There is no question that substantial professional development is necessary for

teachers and leaders about setting appropriate objectives and performance expectations, differentiating instruction, and evaluating objectives in order for an SLO system to be successful. However, the type of professional development required would be very similar to professional development for implementing standards-based instruction. In other words, SLO, more than many other systems for attributing student performance to educators in an evaluation system and if implemented well, should have the collateral benefit of improving teaching and learning directly. More specific to this paper, SLO should be evaluated:

- Using an end-of-course or unit assessment of the highest quality available,
- Without using a pretest-posttest design, and
- With the most appropriate analytic method available and/or using a "rough conditional status model."

We elaborate on these recommendations here. At a very simple level, given the many challenges of trying to find or develop two high quality assessments for each course every year, it just makes much more sense to base the evaluation of the SLO on a high quality end of course or end of unit assessment. Now the problems with simple status models are well known in that they are highly correlated with out-of-school factors such a family background and wealth, so we further recommend that the status evaluation be conditioned on available prior scores or other information to help make the system more fair. If the data are appropriate, the sample size large enough, and the analytic capacity available, a regression-based approach (including ANCOVA) could serve as the analytic frame for the SLO.

However, it unlikely that such analytic capacity will be found in most districts. Therefore, the conditioning should be much courser grained whereby the students are placed into one of three or four initial starting groups based on past achievement data and other relevant information. The "objectives" would then be differentiated according to the students' starting group. Of course there is the risk that certain students could be targeted with lower than appropriate expectations, but auditing systems, including perhaps the use of external prior scores, should be developed to prevent against such unintended consequences.

Of course, this makes clear that there is no perfect solution to the problem of evaluating SLO, but a "conditional status" approach appears to reduce many of the risks of trying to implement a growth-based system, while offering several potential benefits. Clearly research is needed to address these questions empirically, but the short term focus, given the consequences of these evaluation systems, should be on minimizing potential unintended negative questions. Any SLO system as well as the larger educator evaluation system must be piloted and subjected to rigorous evaluation so that we can learn about the most useful and valid approaches for implementing SLO.

Again, the focus in this paper has been on SLO. Many of the concerns discussed in this paper, particularly about assessment quality and weak analytic approaches, are not limited to SLO. In

fact, the "test-based" approaches being used in many places like Hillsborough, FL or Harrison, CO (Buckley & Marion, 2011) are even more susceptible than SLO to assessment quality and analytic capacity shortcomings. SLO have the potential advantage of being more robust than more straightforward test-based system, where data in addition to the assessment results, can be brought into the decision making process. With test-based systems, the validity of the inferences rests much more heavily on the quality of the assessment and the appropriateness of the analytic methods than with SLO. Again, SLO do not have free pass here, but as noted previously (Marion & Buckley, 2011), they offer the most promising way forward at this point.

References:

Braun, H., Chudowsky, N., & Koenig, J. A. (2010). *Getting value out of value-added: Report of a workshop.* Washington, DC: National Academies Press.

Briggs, D. C. (in press). Making Value-Added inferences from large-scale assessments. Simon, M., Ercikan, K., & Rousseau, M (Eds.) *Improving Large-Scale Assessment in Education: Theory, Issues and Practice*. London: Routledge.

Buckley, K. & Marion, S.F. (2011). A Survey of Approaches Used to Evaluate Educators in Non-Tested Grades and Subjects. Retrieved April 1, 2012, from www.nciea.org.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66, 3*, 227-268.

Diaz-Bilello, E.K. (2011). A validity study of interim assessments in an urban school district. Unpublished doctoral dissertation. University of Colorado, Boulder.

Goe, L. & Holdheide, L. (2011, March). Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects. Princeton, NJ: National Comprehensive Center for Teacher Quality.

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R.L. & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis,* 29-36.

Marion, S. F. (2009). Changes in assessments and assessment systems over the past decade. Presentation to the National Research Council's Workshop on Best Practices in State Assessment. Washington, DC.

Marion, S.F. & Buckley, K. (2011). Approaches and considerations for incorporating student performance results from "Non-Tested" grades and subjects into educator effectiveness determinations. Retrieved March 2, 2012, from www.nciea.org.

Perie, M., Marion, S.F., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28, 3,* 5-13.

Rothstein, Jesse. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537-571.

Secretary's Priorities for Discretionary Grant Programs, 75 Fed. Reg. 47,288 (proposed Aug. 5, 2010). Retrieved April 2, 2012, from http://www2.ed.gov/legislation/FedRegister/other/2010-3/080510d.pdf

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29, 7,* 4-14.

Wiggins, G. & McTighe, J. (1998). Understanding by design. Alexandria, VA: Association for Supervision and Curriculum Development (ASCD).