THE COLORADO EDUCATION INITIATIVE

APPROACHES AND CONSIDERATIONS FOR INCORPORATING STUDENT PERFORMANCE RESULTS FROM "NON-TESTED" GRADES AND SUBJECTS INTO EDUCATOR EFFECTIVENESS DETERMINATIONS

Scott Marion, National Center for the Improvement of Educational Assessment
Katie Buckley, Harvard University
September 7, 2011

APPROACHES AND CONSIDERATIONS FOR INCORPORATING STUDENT PERFORMANCE RESULTS FROM "NON-TESTED" GRADES AND SUBJECTS INTO EDUCATOR EFFECTIVENESS DETERMINATIONS

Scott Marion[1]

National Center for the Improvement of Educational Assessment

&

Katie Buckley

Harvard University

September 7, 2011

[1] We are grateful to the Bill & Melinda Gates Foundation for supporting the production of this paper, which is one of a series of papers on "non-tested subjects and grades." The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Foundation.

Approaches and considerations for Incorporating Student Performance Results from "Non-Tested" Grades and Subjects into Educator Effectiveness Determinations

There has been a growing interest in reforming the long-standing negotiated approaches for evaluating and compensating teachers by among other things incorporating student performance results in teacher evaluations. Advances in growth and value-added models in education have contributed to the interest in using changes in student test scores over time as part of educator accountability systems (Baker, et al., 2010; Braun, et al., 2010). Many districts, states, and non-governmental organizations (e.g., New Teachers Project) have embraced these test-based accountability initiatives, but the initial focus has been on the content areas and grade levels for which there are state standardized tests, generally administered at the end of each school year. Education leaders, especially those submitting Race-to-the-Top (RTTT) applications, have quickly realized that evaluation systems focusing on teachers in subjects and grades for which there are state test data generally means that only one-quarter or so of the teaching force would participate in such evaluations. As discussed by Buckley and Marion (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 subjects and grades not assessed by state standardized tests.

This paper is designed to help policymakers and accountability professionals wrestle with the challenges of using student performance information as a component of educator evaluations when yearly state standardized tests are not available. We first present a brief overview of potential measurement tools and analytic approaches for non-tested subjects and grades followed

by a discussion of some technical challenges inherent in these tools and approaches. Based on this discussion, we offer recommendations for how states may proceed with creating educator effectiveness systems given the technical challenges that exist. We recommend that states apply a theory of action to their educator effectiveness system to illuminate those approaches that might be fraught with the greatest challenges and those that might have the most potential in measuring educator effectiveness, and consider broadening student learning objectives as a framework for incorporating student performance information in educator evaluations.

Measurement Approaches for Determining Educator Effectiveness

The data available for calculating student growth in grades and subjects other than grades 4-8 reading and mathematics is quite variable and often very limited. One of the challenges in thinking about measures for use in educator evaluations is that these considerations are intricately connected to the multitude of teaching responsibilities. Eighth grade science teachers (with a state test at the end of 8^{th} grade) will have different data available than a high school art teacher. Therefore, one of the first tasks when considering the measurement approaches for non-tested grades is to first be clear about the teaching responsibilities and the data sources associated with the specific teaching responsibility that may be used in an evaluation system. Appendix A provides an example for how states may identify the measurement tools and available data associated with various groups of teachers as well as the responsibilities the various educators have for directly influencing student achievement.

The second major consideration is to define the required levels of technical quality required before assessment results are used in educator evaluations[2]. There is considerable tension in determining the expected degree of technical quality. Most current state assessments meet minimal levels of traditional notions of technical quality, but fall short of more progressive conceptions of technical quality such as measuring the expected depth of student learning intended by most state standards and especially the Common Core State Standards. However, if we impose such a high bar for the technical quality of assessments used for educator evaluation, it is doubtful that anything other than a state standardized assessment will meet these criteria. **On the other hand, if these assessments are used to make important decisions about the placement and compensation of educators, they should be held to a high bar or their use should be limited until the technical quality can be assured**. We briefly discuss several of the predominant approaches/measurement tools for evaluating teachers who teach in non-tested grades and subjects. We must be absolutely clear about one thing: There is no single solution to this challenge; rather a comprehensive strategy is required to address the full set of non-tested grades and subjects.

Creating and/or Selecting New Tests

A common approach (at least initially) for helping to close the non-tested gap is to create more state-level or at least state- or large district-sponsored tests in subjects and grades where there are not current large-scale tests. States such as Delaware and large districts such as Hillsborough, FL and New York City are pursuing, to varying degrees, creating new tests. This is an

[2] See Buckley & Marion (in progress)

intuitively appealing approach to address a difficult problem. Unfortunately, while creating a testing program is difficult, maintaining high quality testing programs is even more challenging.

Using tests for high stakes purposes - and no one can question that educator accountability involves high stakes - requires that the tests meet very high standards of technical quality, including a level of reliability necessary to support high stakes decisions, items and forms that meet content validity standards, and technically appropriate linking designs to ensure that scores across years or forms can be validly placed on the same scale. This last point is critical, because in many of the discussions about creating new tests, much of the rhetoric makes it sound like this is a one-time effort. In fact, when tests are used for high stakes purposes as part of student growth calculations, security considerations require that new forms be created for each new use. Being able to link these forms across years or uses requires sophisticated linking (equating) designs. Most states and school districts do not have the internal psychometric capacity to support such efforts and this would require contracting with a test development company or consultant. Therefore, the ongoing costs of such an initiative could be significant.

We are not advocating a massive effort to create new tests because it is likely that the return on investment for such an effort would be quite low. In other words, after spending significant amounts of money to build new tests, states and districts would still be a very long way from making a major dent in the problem. That being said, we are not opposed to creating new assessments in potentially high leverage situations, such as where a state assessment exists in one grade and there is no prior score to use for calculating growth. Further, creating new

assessments in strategic areas could help support a higher quality implementation of Student Learning Objectives (discussed below) than might otherwise be the case.

Those tests could provide data for calculating growth or value-added quantities. Not that doing so would alleviate all concerns with using test scores to evaluate educators (e.g., Baker, et al., 2010, Braun, et al., 2010, Briggs & Domingue, 2011)[3]. More realistically, though, the resources are simply not available to support such an endeavor and few people would argue for such an increase in external testing in the first place. Even if resources were available, students do not typically take courses in a common sequence, especially in secondary school, so that adding more tests might not provide the solution that many hope it would because large-scale VAM or growth models generally work best when students have multiple prior scores from a common course sequence. On the other hand, if the tests from a particularly content domain are all fairly well correlated, it might not matter from a statistical point of view whether students took courses in different sequences, but it could matter very much when trying to ensure the public credibility of such an approach. One way that districts and states are looking to avoid the course sequence concerns, especially in high schools, is to employ a pretest, posttest design, whereby the pretest[4] would be administered early in the school year and the posttest would be administered later in the same school year. Of course, using this approach means at least twice as much testing as the year-to-year approach, close to twice the cost, and involves a loss of twice as much instructional

[3] For example, many critics have challenged the validity of VAM models as a measure of teacher effectiveness by documenting the unreliability of teacher rankings based on VAM scores for a given year (Baker, et al., 2011, Braun, 2010), the way in which particular model choices influence interpretations of effectiveness (Briggs & Domingue, 2011), and the challenges of attributing the student growth to the appropriate educators (Baker, et al., 2011, Braun, 2010).

[4] There are several possible pretest designs that need to be considered carefully. These include a test that is essentially a parallel test to the end of year test, a measure of key precursor knowledge and skills required for success in the course, or some combination of the two as well as other potential designs.

time. Those potential negative consequences pale in comparison to the risk of cheating and other forms of score corruption[5].

[5] With a pretest/posttest design, teacher have a strong interest in making sure that pretests are as low as possible, while posttests are as high as possible. In a spring-to-spring design, each teacher is interested in seeing their students score as well as possible.

Interim Assessments

Others have suggested that interim (or benchmark) tests could supplement current state tests or fill the need in these non-tested subjects and grades. In fact, this suggestion was found in the RTTT application materials and not surprisingly was found in many of the applications submitted by states (Buckley & Marion, 2011). It might be possible to have some interim assessments used in subjects and grades where the technical quality and intended uses are appropriate, but there are several problems with this hope. First, as many have documented, the technical quality of current interim assessments leaves a lot to be desired (e.g., Bulkley, et al., 2010, Goren, 2010, Li, et al., 2010, Perie, et al, 2009, Shepard, 2010). Inserting interim assessments into this gap might solve one problem. On the other hand, such an approach would create many more problems than it would solve. Based on the low quality of the current crop of commercially-available interim assessments (especially in terms of item quality and rigor), their increased use could send a contradictory message about the meaning of college and career readiness as well as running the risk of narrowing the curriculum in ways that do not support the current RTTT and CCSS reforms. Further, even if these assessments could be used to fill some of the need in these non-tested grades and subjects, a significant gap would remain because most interim assessments have been targeted to reading and mathematics at the exclusion of other subjects as a way to help schools prepare for NCLB end of year tests (Perie, et al., 2009). Some of the interim assessment vendors may offer "science" or "social studies" assessments in selected grades, but at best these are only reasoning tests that include very little science or social studies content. Even if some of these interim tests were worth using, policy makers still would be left with the problem of having many teachers in grades and subjects without any external tests.

Classroom and related assessments

Several states and entities have proposed a variety of approaches that revolve around using specific aspects of the classroom-based assessment system as a means for determining how much students have learned either through a particular school year or from one year to the next. Some approaches involve simply feeding the data from these classroom-based measures into some sort of analytic method used to calculate growth or a value-added quantity[6]. Others have proposed a variation on this theme whereby teachers use classroom-based and/or other information to establish goals for either individual students and/or the class as a whole and then evaluate the degree of success in terms of meeting these goals using similar or other relevant measures. Each approach addresses certain issues while raising different challenges. Common to both approaches is the desire on the part of policy makers and educational leaders to continue to use these assessments as part of the teaching and learning cycle, thereby providing instructional feedback, in addition to their use in educator effectiveness determinations. An obvious advantage to this approach is that the amount of additional, external testing is limited. Unfortunately, there are many significant challenges when trying to use assessments for both high stakes accountability and for any other educationally helpful purposes. Creating and utilizing a theory of action will help reveal these tensions and will likely lead the designers to recognize that they have to prioritize one purpose over others.

Methods to Analyze Student "Growth"

Much of the discussion around incorporating student performance information into educator evaluations has focused on the "what" or the assessments (tools) that will be used to measure student performance. This has certainly been appropriate as a first step. However, once school

[6] We discuss the challenges and complexities involved in applying VAM to classroom measures later in the paper.

and district leaders determine these tools, the next step is the critical one of "how." There are many different methods for attributing student test scores to teachers, but simply having two scores for each student (e.g., pretest, posttest) does not automatically imply a method for evaluating these scores. There are many methodological choices that must be considered when determining how to most validly analyze and incorporate student performance information in educator evaluations. This section describes several of the most commonly used "families" of methods for documenting "student growth" while outlining some of the technical and practical issues associated with each approach, particularly as applied to the non-tested subjects and grades context.

Growth Models

In many ways, growth models are the holy grail of student longitudinal modeling. It is actually what most users would really like to know—how much more did this student learn this year compared to last? True growth models require ordered content expectations and tests designed using vertical score scales that have interval scale[7] properties. Unfortunately, content expectations are rarely ordered in ways that would permit such interpretations, especially once students progress past third grade. Further, tests are rarely designed in ways to permit equal interval interpretations, especially across grades[8], where changes in content knowledge are perfectly related to change in scores. Nevertheless, growth models are still used in many contexts, although validity evidence to support such scales may not be persuasive. Growth

[7] Interval scales are ones where the difference in scores at one point on the same scale have the same meaning as the same nominal differences on another part of the scale (e.g., the difference between a scale score of 220 and 240 on a test means the same thing in the context of student growth as a difference in scores from 240 to 260).

[8] While some advocate the use of vertical score scales to address this concern, the research is quite clear that vertical scales used for current state tests do not meet these equal interval assumptions. Many would go further and say that they don't even come close.

models require at least two scores in an ordered domain (to the extent possible) and must be on the same scale. Ideally the scale should have interval properties or at least properties that approximate an interval scale. If the assessments are not designed to be on the same scale, the scores should be transformed such that the scores might be compared in valid ways. Finally, in and of themselves, growth scores do not say anything about teacher effectiveness, and policymakers must decide how to use results from the growth models to say something about the influence of the teacher(s) on student growth.

Value-added models

The following three approaches described here—VAM, Conditional Status, and SGP—are all being considered by states in order to measure teacher effectiveness. These methods tend to be operationalized in similar ways and operate from similar principles in that they all use some type of "prior" assessment score and perhaps other factors to condition ("adjust") the posttest results. These approaches provide a way to characterize the performance of students n a grade/subject by relying on a prior test and current (post) test in a given subject. Unlike a true growth model, these approaches do not provide a measure of student growth, but provide a relative measure of change in student achievement by comparing and ranking (implicitly or explicitly) student achievement gains among students with similar characteristics (i.e. prior test scores). VAM models are the most well known of these complex analytic methods, but in reality, there is no single VAM, rather a class of models that have intuitive appeal because they appear to "level the playing field" with the hopes that the results can be treated as if they are a result of a clear cause and effect link (e.g., student learning results produced by the model can be attributed to the teacher). This can

never be the case since students are not randomly assigned to teachers and schools (see Rothstein, 2009), but that does not appear to reduce the intuitive appeal.

Value-added scores are generally derived from regression-based or ANOVA-based models, and require at least two test scores (although additional years may be included to improve the precision of the estimates), and may include additional covariates such as student demographics or school characteristics. VAM scores are interpreted as the difference between a student's predicted score (based on similar students) and actual score; a difference that is attributed to the teacher in a VAM. If a student's observed score is greater than their expected score, indicating that the student performed better than would be expected based on the performance of other similar students, the difference is positively attributed to their teacher. Since a key piece of such models is the need for pre- and post-test data for each student (e.g., a test in third grade and then in fourth grade covering the same domain or a pretest and posttest in the same grade such as fall and spring), a potential solution for non-tested grades and subjects is to create or implement new tests in those areas, such as those mentioned above.

Conditional status models

When pre-test data in the same subject does not exist, some states are exploring the approach that we have termed the "conditional status" approach, however these are really just value-added models without prior scores from the *same* subject area. States with end of the year assessments for certain grades, as is the case for many grades that administer AP exams and current NCLB science assessments typically administered in only three grades, are considering this approach. This model uses students' earlier scores in another subject to statistically control for current

performance on a summative assessment. For example, if no true pre-test exists for an 8th grade science test, states may include a standardized measure of prior math or reading scores (or both) as a control variables in a model, and measure achievement among students in 8th grade science conditioned on prior math/reading test scores. Like VAMs, "teacher effectiveness" is determined by comparing observed changes in learning to predicted changes in learning, and attributing the difference to the teacher.

Student growth percentiles

While student growth percentiles were developed primarily for descriptive purposes, the mathematics underlying SGP are fairly similar to many VAM approaches and have been applied in accountability models. Student growth percentiles are calculated using quantile regression techniques whereby students with the same score history are compared with one another and their relative position on the posttest (current test) is described using a percentile metric. On a statewide level[9] , the "average" score will be, by definition, the 50th percentile (median). Therefore, students with percentiles greater than 50 performed better than their peers, while those with percentiles less than 50 performed worse than their academic peers.

Student growth percentiles can be aggregated to any unit desired, and therefore if aggregated at the classroom level, they have been used to derive a measure of "teacher effectiveness". At the classroom level, the median growth percentile can be used to describe the growth of students associated with a particular teacher. Classrooms with median growth percentiles greater than 50

[9] The unit of analysis may be at the district or even school level as long as the sample size is large enough. For example, SGPs can be calculated at the district level, assuming the district is large enough (e.g., 1000 students per grade level), but it will be important to remember that, by definition, the district median SGP will be 50, so all student and aggregate results will be compared to the district and not state average.

had students that, on average, performed better their peers. Many of the data requirements for SGPs are similar to VAM, except that SGPs have typically been calculated using statewide (as opposed to district) test scores as a way to better contextualize the interpretations

School-wide attribution

In the absence of strong pre-and post-test data on newly implemented tests, some schools and districts are attributing school-wide gains from the state assessment to individual teachers. This approach has been proposed by at least Tennessee and Maryland, as noted in Buckley and Marion (2011) and is encouraged in Colorado's system. Often times, school-wide attribution makes use of traditional value-added models and simply generates a "school effect" instead of a "teacher effect". Of course, with this approach, student learning on state assessed subjects—typically mathematics and ELA—is attributed to all teachers including those who teach subjects unrelated to these fields. This approach is thought to encourage school-wide collaboration, but others worry that it reduces variability so much that it does not allow for determining effective from ineffective teachers.

Student Learning (Growth) Objectives

Finally, many states are considering the use of student learning (or growth) objectives (SLOs) for grades and subjects where implementing a standardized assessment is infeasible. With the SLO approach, teachers use classroom-based and/or other information to establish goals for either individual students and/or the class as a whole and then evaluate the degree of success in terms of meeting these goals using similar or other relevant measures. The RTTT Technical Assistance (TA) Network defines SLOs as "a participatory method of setting measurable goals,

or objectives, based on the specific assignment or class, such as the students taught, the subject matter taught, the baseline performance of the students, and the measurable gain in student performance during the course of instruction" (2010). In providing our recommendations below, we argue that SLOs can be broadened to serve as an overarching framework for evaluating teachers.

Technical Considerations of Measures and Methods

As policymakers make decisions about which measurement tools and analytic methods to use, they must pay attention to the various data and technical requirements of each approach. In a forthcoming paper, we outline some specific technical requirements for each of the various assessments and analytic techniques. We highlight some of the key principles undergirding these requirements below. These requirements need to be built out in the form of both guidance and support materials for districts. While many technical requirements are specific to each of the particular techniques, the following general technical principles apply to all of the student "growth" methods:

- Assessments should be technically adequate to support the intended analyses,
- Analyses shall be based on a large enough number of students to warrant reasonably consistent inferences,
- The particular approach (model) should make design choices explicit and transparent and where sufficient technical documentation exists to judge the technical quality of the approach (this is especially true for techniques such as SGP, VAM, etc), and
- The model (or those implementing the model) should produce results in ways with the greatest likelihood of effective use.

In the sections that follow, we outline criteria specific to each of the approaches. Clearly, there is significant overlap in the criteria for many of the general approaches, but there are also important factors unique to each approach.

Growth models

1. Pre and post test scores from assessments in the same subject with student-level correlations of at least 0.6 or better
2. Assessments should be technically adequate to support the intended analyses such as:
    a. Both assessments must meet minimum reliability thresholds (e.g., 0.8).
    b. Both assessments should be aligned to the same content domain in conceptually coherent ways such that the assessment scores are thought to have a common meaning across tests,
    c. The assessments need to be on a common scale of some type. If the assessments are not designed such that scores are based on a common underlying scale, the scores from each assessment will need to be transformed in a technically defensible manner (e.g., z-scores) so that scores can be compared appropriately (note: percent correct is NOT a common scale), and
    d. Each test has sufficient "stretch" or variability in the scores to avoid ceiling and floor effects.
3. The model should be evaluated such that biases can be identified and steps can be taken to ameliorate them. For example, if higher achieving students tend to exhibit higher growth (regardless of teacher), then classes with more high achieving students will produce higher "teacher growth scores" than those classes with lower achieving

students. This can be particularly problematic with simple growth models, because these models typically do not include statistical adjustments or controls for pre-existing differences (as is the case for VAM or SGP models) such as demographic characteristics or prior scores except in the case that such scores are used for subtraction purposes.

4. The model (or those implementing the model) should produce results in ways with the greatest likelihood of effective use.

Value-added models

1. Pre and post test scores from assessments in the same subject with student-level correlations of at least 0.5 or better
2. Sample sizes need to be robust in terms of the number of teachers and students. Ideally this means at least 20 students per teacher and, depending on the number of variables included in the equation, at least dozens of teachers in the analyses [I actually think this should be larger, but will appreciate the panel's input]
3. Assessments should be technically adequate to support the intended analyses such as:
    a. Both assessments meet minimum reliability thresholds (e.g., 0.8),
    b. Both assessments should be aligned to the same content domain in conceptually coherent ways such that the assessment scores are thought to have a common meaning across tests,
    c. Each use a scale that at least approximates interval properties (note: a vertical scale is not required), and
    d. Each test has sufficient "stretch" or variability in the scores

4. The model should be one where design choices are explicit and transparent and where sufficient technical documentation exists to judge the technical quality of the model
5. The model (or those implementing the model) should produce results in ways with the greatest likelihood of effective use.

Student growth percentiles

1. Pre and post test scores from assessments in the same subject with student-level correlations of at least 0.5 or better (multiple prior scores are preferred)
2. Analyses are most meaningful when performed on statewide samples, but could also be conducted within-district for large districts (e.g.,, at least 1000 students per grade/test). Note that interpretations for district analyses are all centered on a district average compared with the state median when used for state analyses. Sample sizes generally need to be at least 20 students per teacher.
3. Assessments should be technically adequate to support the intended analyses. These criteria need to be more fully developed, but should include at least the following:
    a. Both assessments meet minimum reliability thresholds (e.g., 0.8),
    b. Both assessments should be aligned to the same content domain in conceptually coherent ways such that the assessment scores are thought to have a common meaning across tests,
    c. Each use a scale that at least approximates interval properties (note: a vertical scale is not required), and
    d. Each test has sufficient "stretch" or variability in the scores (e.g., limited ceiling effects)

4. The model should be one where design choices are explicit and transparent and where sufficient technical documentation exists to judge the technical quality of the model
5. The model (or those implementing the model) should produce results in ways with the greatest likelihood of effective use.

Conditional status models

1. Pretest scores should be correlated with posttest scores (at least 0.5 or preferably better) and come from a subject area where a case can be made that the pretest and posttest scores are at least conceptually related (e.g., reading test serving as a pretest for a social studies test)
2. Sample sizes need to be robust in terms of the number of students per teacher. A typical rule of thumb in regression analyses is to have 15 cases per variable, which would mean that this type of analysis would require a minimum of 15 students per teacher and closer to 30 would be better.
3. Assessments should be technically adequate to support the intended analyses such as:
    a. Both assessments meet minimum reliability thresholds (e.g., 0.8),
    b. Each use a scale that at least approximates interval properties (note: a vertical scale is not required), and
    c. Each test has sufficient "stretch" or variability in the scores
4. The model must be evaluated for bias to ensure that analyses do not produce over- or under-predictions, especially differentially for specific subgroups of students.
5. The model should be one where design choices are explicit and transparent and where sufficient technical documentation exists to judge the technical quality of the model.

6. The model (or those implementing the model) should produce results in ways with the greatest likelihood of effective use.

Attributing school-wide growth on a state summative assessment to individual teachers

1. There are very few technical considerations when using this approach, but at a policy/values level, decision about attribution need to reflect school values and employ a shared decision-making approach for determining levels and types of attribution.

Student growth objectives

1. Each district shall develop a set of procedures for establishing and evaluating goals. These procedures shall include general district approaches as well as providing guidance for specific content areas.
2. Goals shall be established for each student and at the aggregate classroom level, such that individual students are ambitious and standards-based, while aggregate goals may be normative. We strongly suggest having aggregate goals focus on the full range of students rather than the simple class average.
3. Goals shall be based on data such as prior assessment/grades history and must reflect meaningful (e.g., college readiness) and measureable targets.
4. Multiple goals may be established for each student, but at least one of the goals shall be a long-term goal (e.g., a semester or year) in order to have a greater chance of detecting real change.

5. Goals shall be set by teachers in consultation with professional learning communities, a committee of peers, and/or principals. Goals should be made public, at least internally to other educators in the school and parents.
6. Progress toward and attainment of goals shall be determined by measures that are aligned with the learning targets and are technically appropriate to determine whether students have actually met the goals. In other words, a case in which the assessment should be avoided is if it is only least nominally aligned with the targets and at a level far below the actual goals so that one is unable to actually judge if the student met the goal.
7. The assessments used to measure the goals shall be reviewed by a committee of peers and administrators to judge their adequacy for evaluating student progress towards the goals.

The criteria outlined above represent initial thinking in this area. A current project is further exploring and developing criteria and guidance that states may use to evaluate assessments and methods for incorporating student performance information in educator evaluations.

Recommendations

We offer three primary recommendations as states move forward with building their assessment and accountability system for evaluating teacher effectiveness. The first is to develop an explicit theory of action to illuminate how the system is intended to work and to identify potential unintended negative consequences. The second is to consider using SLOs as an overarching framework for incorporating student performance into educator accountability systems. The third recommendation is based on the field's limited knowledge about "best practices" and we offer

suggests for learning from early implementation efforts. We discuss these recommendations below.

Developing a Theory of Action for Accountability Decisions

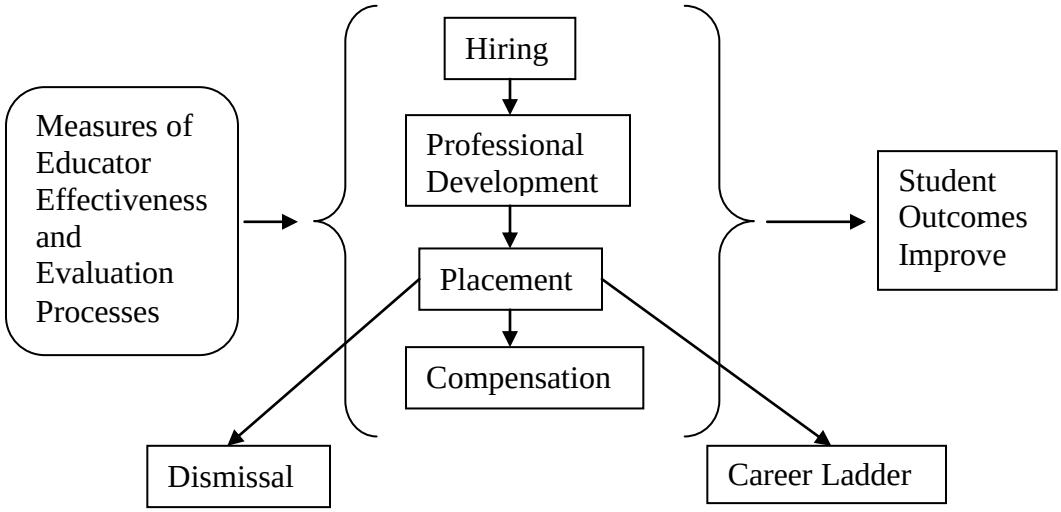
One of the axioms in validity theory is that an assessment can be validated only within the context of a specific purpose and use. Yet, the multiple demands being placed on assessments proposed under both the Race to the Top accountability proposals and the Race to the Top Assessment program (RTTA) require a renewed attention on the use context in the evaluation of technical quality of these assessments and assessment systems. Developing and using a theory of action to guide design and evaluation can help illuminate policy alternatives and potential competing or even contradictory claims about how assessments will function in an educator evaluation system. Policy makers, developers, stakeholders, and technicians must very explicitly lay out why implementing test-based approaches to support educator effectiveness for these grades and subjects will lead to improved educational opportunities for students. In addition to the why, policy makers should have to describe the how, or the mechanisms by which they think that these improved learning opportunities will occur. For example, one might postulate that holding teachers accountable for increases in student test scores on classroom-based assessments will lead to the development of both better assessments and improvements in student learning. The evaluator and/or user must specify the mechanism by which these accountability uses will lead to the anticipated changes in teaching practices, such as having teachers better differentiate instruction to students and/or the development and use of more appropriate curricular materials. **We argue that articulating the aims and mechanisms of the program via a theory of action will expose many of the proposed policies for evaluating educators in non-tested (as well as**

tested) grades and subjects as untenable, but will also shed light on some fruitful means of meeting the major policy goals.

The Overarching Theory of Action

It is helpful to step back and present a big-picture theory of action for the entire educator evaluation policy. This is presented to help contextualize the current discussion, but to also serve as an example of a theory of action[10].

Figure 1. A theory of action for educator evaluation.



This theory of action illustrates that an educator evaluation system along with measures of effective teaching are expected to influence hiring decisions (assuming we could use such evaluation approaches with pre-service educators), professional development planning, placement, and compensation decision. Obviously, this assumes coherent links between the educator evaluation results and these decisions. Ultimately, these steps and mechanisms are intended to lead to either dismal for ineffective educators or a career ladder for varying degrees

[10] We are grateful to Brian Stecher for this example.

of effective educators. Ultimately, these various processes and mechanisms are hoped to lead to improved student outcomes. Now, it is one thing to draw a nice neat picture and another thing to posit a theory of action. Theories of action must be testable (falsifiable in a scientific sense) in that each of the claims implied by the specific components and connections among components must be made explicit. In the case of this very simple theory of action, one would want to search for evidence that the evaluation measures influenced hiring, placement, compensation, and professional development in positive ways. To the extent that the evaluation measures negatively influenced these things (e.g., hire quality teachers were seen leaving the profession at higher rates than low quality teachers), it could count as a threat to the validity of the system. Ultimately, the entire system is intended to improve student outcomes. These sorts of distal outcomes are very difficult to connect, in any sort of causal way, to specific policies and programs. Therefore, it is often helpful to include some sort of intermediate variable, such as "teacher practices improve" between the set of program/policy variables and the "student outcomes" so that evaluators and others do not have to wait multiple years to begin judging the effectiveness of the policy. As seen in this abbreviated example, theories of action are a useful tool for explicating the expected workings of a policy or program to both allow for formative monitoring of the implementation and to serve as a basis for validity evaluation.

Theories of Action to Illustrate Components and Mechanisms

Figures 2 and 3 attempts to explicate the tension between using assessments for instructional and accountability information by showing the differences in components and mechanisms required for the tests to support the differing uses. This first set of representations assumes at least one pretest (prior score) and at least one posttest. A limitation of these graphical representations is

that they cannot reveal all of the important mechanisms or processes that are the critical connections among the highlighted components. We show that it is the differences in these mechanisms that make the two theories of action irreconcilable. Further, these are highly simplified theories of action, but we are using these to illustrate some key principles.

As can be seen in Figures 2 and 3, the expectations for the pretest are different, depending on the intended use. In the instructional case, it is important that the pretest measure either the expected prior knowledge or the forthcoming required knowledge and skills well enough so that the tests provide insights for the teacher as she plans her upcoming instructional activities. This is depicted on the left side of Figure 2 and the mechanism label #1 in this figure. The accountability use for such a pretest simply requires that the pretest be correlated with the posttest, but in reality it should be conceptually related to the posttest to make for a more valid accountability determination. This is one of the areas of the theories of action where the differences in use would prevent using the richer pretest used in the instructional setting for accountability as well, at least on the surface.

The mechanism labeled #4 in Figure 3, the accountability theory of action represents an important choice point for many educators. Most often, this is not an explicit choice because many have either been led to believe or naively believe that focusing on improving test scores is the same thing as ensuring that students are learning the knowledge and skills that actually represent the intended domain (generally defined by the standards or curriculum). This is one of the major concerns when assessments intended to support instructional uses are shifted to accountability uses. Teachers may focus as expected, depending on the stakes associated with

the accountability determinations, on increasing test scores instead of ensuring that students are learning the content at the intended depth. There is long literature documenting this unfortunate choice of actions, especially as it affects poor and minority students. On the other hand, when tests are designed to support instruction, the test (and reports) must yield information relevant to instruction and teachers must be able to interpret this information and know what to do next in terms of instruction (see Figure 2, mechanism #2). Neither of these components of the instructional theory of action should be taken for granted because both would require compelling evidence to support claims that assert, for example, that teachers possess pedagogical content knowledge adequate for being able to properly interpret student performance information and make appropriate instructional adjustments.

The next component of the instructional theory of action expects that the prior actions will lead to improved student learning (mechanism #3), whereas the accountability theory of action assumes only that students test scores will increase compared to the pretest (mechanism #5). These increased test scores should lead to higher accountability results which are intended to motivate continued and improved actions by teachers. On the instructional side, we would expect the improved student learning to lead to higher posttest scores.

Reconciling the different uses and intentions

The close examination of these two theories of action makes it appear that they are too divergent to find any reconciliation. Given the intense interest among many policy makers as well as local education personnel in not adding any additional external assessments to the current mix of state and local assessments, while maintaining the possibility of using student outcomes from these

"non-tested" courses in educator evaluations, it is worth taking a closer look to search for possible reconciliation. In this case, I argue that the instructional theory of action should be considered the primary use since that is the current use case, but we will work to see if accountability uses can be accommodated.

On the surface, a case could be made that the evaluation of the change in performance between the pretest and posttest scores for the entire course would not disrupt the instructional purposes, but would still allow for accountability determinations. In a perfect, non-corruptible world, this might be true, but let's examine some of the measurement issues before getting into the shortcomings of human beings.

A test used early in the school year for instructional purposes would probably try to determine students' competence knowledge and skills judged to be important in successfully completing the requirements for the current course. For example, a high school chemistry pretest for instructional purposes should include enough items to judge students' understanding of proportional reasoning (a critical math skill for high school chemistry), among other topics, to see what type of early mathematics practice/remediation the teacher might need to do with specific students or the class as a whole. However, while these skills will be woven throughout most chemistry curricular, they will not be the focus of the class. Therefore, thinking about a measure of proportional reasoning as a pretest in an accountability context might be hard to defend on content validity grounds in spite of the fact that it might "work" in a statistical sense because it would likely be correlated with an end of course chemistry test. Therefore, it would take considerable effort to design the types of pretests that would provide useful information for

planning and adjusting instruction and that would be more than simply correlated with the end of year tests. Some sort of mini-version of the end of course test might serve as a fair accountability pretest so that analysts may judge how much students have learned relative to the expected domain after they complete the end of year test. Yet, it is hard to see how having students get a lot of wrong answers because they have not had an opportunity to learn the material would serve instructional purposes. I am using the chemistry example purposefully here, because it might be less challenging to bridge the two sets of uses in a domain organized by a learning progression that extended across multiple grades such as elementary reading.

The major differences occur in the next steps of the theory of action, where in the accountability use, the focus is on raising test scores, while in the instructional case, the focus is on increasing and deepening student learning. Now, one could argue that if teachers in either context focused on increasing student learning, the two uses could be bridged. But this is where the practical realities of accountability testing must be surfaced. There is a long history documenting the effects of accountability testing where practices such as teaching to the test are quite common (e.g., Haertel, 1999, Shepard, 2000), which we have seen exacerbated under the accountability pressures of No Child Left Behind (Medina, 2010). It would be naïve to think that teachers, knowing they are being held accountable for increasing students' scores on specific tests, would not focus their instruction on what they expect to be covered on the accountability test. While many teachers may conflate test scores with student learning, one would hope that in a truly instructional context, teachers would focus on ensuring that students are being provided with opportunity to learn the full breadth and depth of the domain, not just what will be on the test.

Finally, both theories of action point to an outcome measure that many would like to believe could be the same measure for either instructional or accountability purposes. There is no doubt that a carefully designed assessment could possibly serve both uses, but for the sake of argument let's think about how assessments might look different if designed to serve these different uses. Depending on the stakes, the accountability test might place a premium on reliability such that there will be a focus on using enough items, many of which would likely be selected response, to generate high reliability coefficients. On the other hand, an end of course assessment designed to serve in an instructional system, would be designed to provide rich evaluative information and would hopefully be designed to probe the depths of students' understanding. Reliability, while not unimportant, would be less of a design concern than construct validity. Such an assessment would include a rich array of performance or open-ended tasks. As noted above, these examples are purposely exaggerated and this is a component of the differing theories of action, where it might be possible to try to use the same assessment in both cases.

This activity has revealed that while there are several aspects of the different uses, highlighted by the theories of action, where it might be possible to find efficiencies by using the same assessments to serve both purposes. However, it is unlikely that some of the fundamental differences will be satisfied such that a single assessment system can serve both purposes well.

Figure 2. Abbreviated Theory of Action for Instructional Uses of Tests

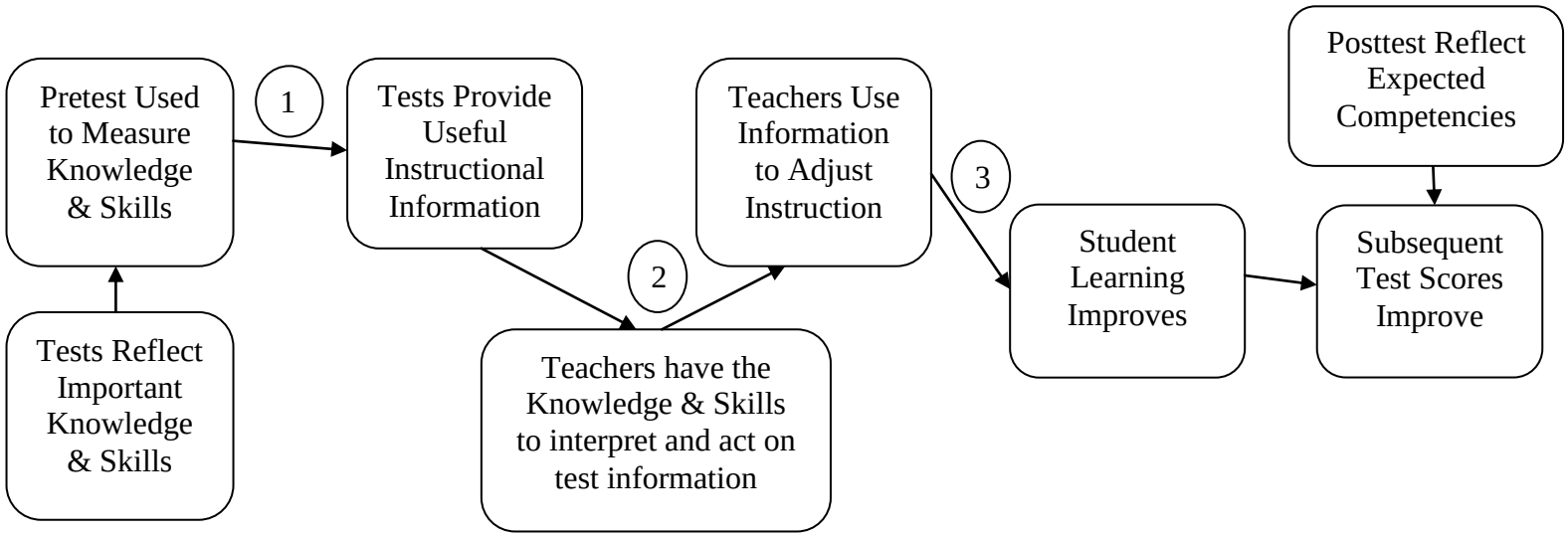


Figure 3. Abbreviated Theory of Action for Accountability Uses of Tests

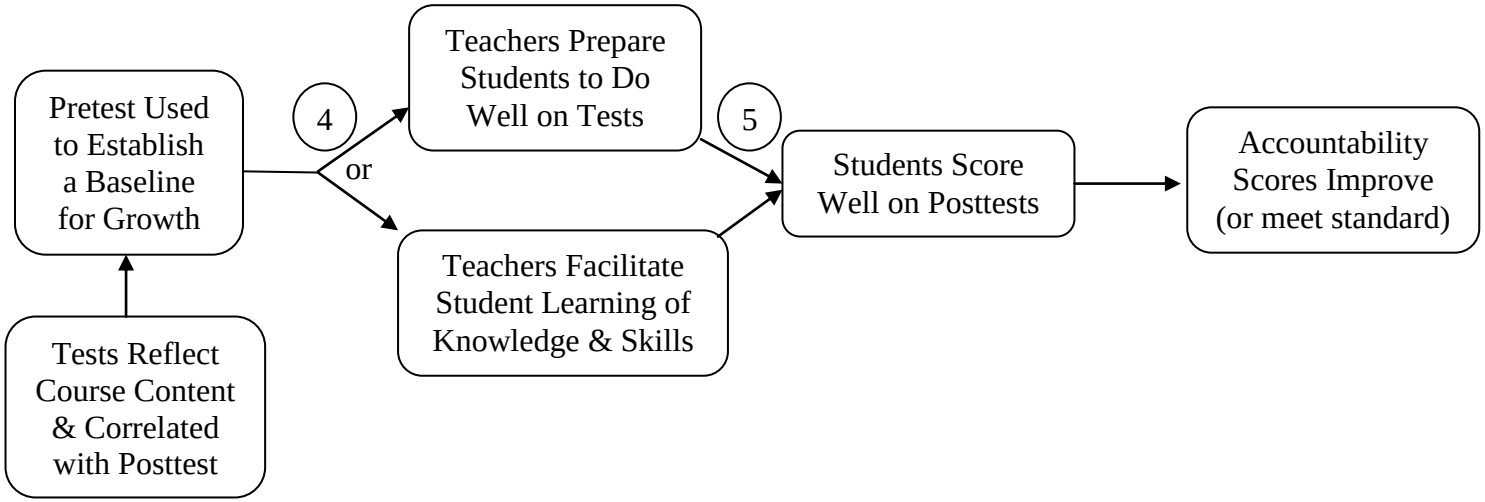
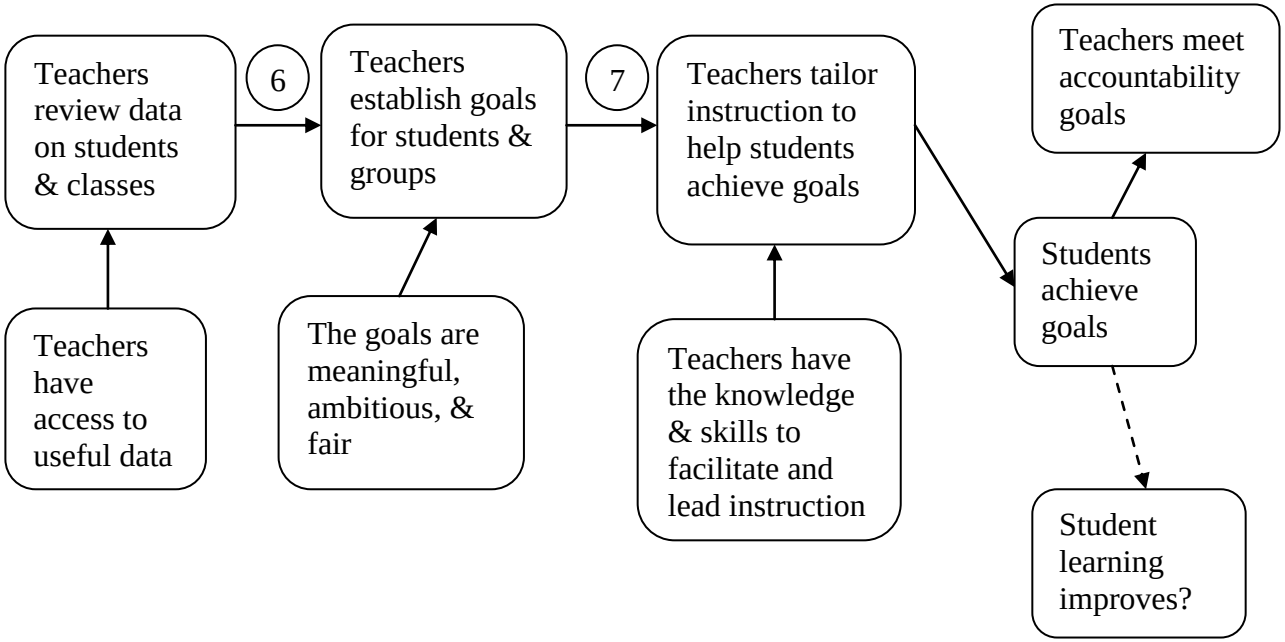


Figure 5. Abbreviated Theory of Action for Goal Setting Approach



Student Learning Objectives as a Comprehensive Framework

Developing a theory of action is a helpful tool to quickly reveal where differences in the use context might cause a conflict when trying to implement a challenging policy. It can also be used to help explore some alternative approaches for satisfying the policy goal of using local assessment information to contribute to the evaluation of educators involved with typically non-tested grades and subjects. Theories of action can (and should) be developed to help evaluate the logic and coherences of some of the alternative policies before moving to the implementation phase. One alternative approach involves trying to use test results from end of year tests from the prior grade to avoid the potential "sandbagging" that could occur when the pretest from the current grade is used as the basis for evaluating teachers. Another alternative is the "goal setting" or Student Learning Objectives (SLO) approach discussed earlier in this paper (see Figure 5).

Student learning objectives offer many advantages over many of the other analytic techniques discussed above. They are highly flexible in that they can be used across all grades and subjects, with existing measures of performance or adapted to new assessment systems as they are developed. Furthermore, SLOs are designed to incentivize the positive practices of setting empirically-based goals for each student (or the class), monitoring the progress toward these goals, and then evaluating the degree to which students met the intended targets. Given this tie to regular practices, it could potentially create greater teacher buy-in to the teacher evaluation system. We do not suggest that SLOs are a panacea to solve the "untested" problem. For example, SLOs can only be as good as the quality of the goals set for each student and by the quality of the measures used to evaluate the goals. Therefore, significant professional learning

opportunities will be required in order to be able to create the learning objectives, ensure that the performance goals set are attainable yet rigorous, and develop or select appropriate measures for the goals.

No matter how many assessments are created in the next few years, the majority of teachers will likely be evaluated using SLOs because of the lack of data available for any other method. As noted above, SLOs are means for promoting positive teaching practices and a way for teachers to internalize important aspects of the evaluation system. As such, we recommend that all teachers use SLOs as part of their student growth determinations. We believe this recommendation will:

- ✓ Provide a common framework for considering measures of student growth for all teachers that can incorporate all other types of types of "student growth" measures,
- ✓ Promote good teaching practices,
- ✓ Foster an internal locus of control in that educators would feel like they have more control over their evaluations compared to externally delivered results,
- ✓ Fulfill a multiple measures goal for all teachers, even those with VAM/SGP results, and
- ✓ Promote a sense of fairness in that all teachers in the school would be operating within the same framework and would all be responsible for designing and evaluating SLOs.

The use of SLOs does not preclude the use of sophisticated techniques such as VAM or SGPs, rather, SLOs provide a context for evaluating the results from such techniques on the context of setting, measuring, and evaluating goals.

Adopting a Continuous Improvement Mindset: Building Evaluation into the System

Including results from student performance, among other factors (e.g., student surveys), in educator evaluations is a very new enterprise. Ongoing, formative evaluations must be put into place alongside new educator evaluation systems as they are piloted and through the first several years of implementation so that we are able to learn what is working well and what needs refinement or reconsideration. Further, the results of these program evaluations, if warranted, must be used to adjust the educator accountability policies and practices. We would make the same case for any new ambitious policy initiative, but this is especially critical in cases such as educator evaluation where we have so little experience and the stakes are potentially high.

As many have pointed out already, there are many technical challenges with using value-added and/or growth measures in educator evaluation system (Baker, et al., 2010, Braun, et al., 2010), but the technical challenges—as we have noted throughout this paper—are even more overwhelming when considering the use of student performance measures in non-tested subjects and grades. States engaged in this work are literally "breaking trail" and it would be quite presumptuous to think that we will get this right out of the gate. Therefore, we strongly recommend that any new approach to educator accountability include a substantial program evaluation component, with significant attention paid to how the system is implemented and is working in non-tested subjects and grades.

A recent report from the Brookings Institution offers a useful framework for conceptualizing how one might evaluate educator evaluation systems, particularly for those educators in non-tested subjects and grades (Glazerman, Goldhaber, Loeb, Raudenbush, Staiger, Whitehurst, &

Croft, 2011). The reported focused primarily on calculating the "reliability" and "power" of educator evaluation systems for the purposes of determining whether these systems can accurately identify "exceptional" educators at either end of the performance distribution. We have concerns with some of the assumptions and orientations in this report, but this is not the forum for offering a full critique of this document. On the other hand, we find the suggestions quite useful for how one might go about evaluating the efficacy of the evaluation approaches used in non-tested subjects and grades (Glazerman, et al., 2011). The authors argue that teachers' aggregate value-added scores (for their classroom or classrooms) in subsequent years is an appropriate criterion for evaluating the capacity of educator evaluation systems (and components of such systems) to "predict" teacher effectiveness. Using subsequent value-added results as the criterion can certainly be questioned, but we think the framework is useful. Of course, the reader will surely notice the problem with this framework for evaluating the quality of evaluation systems for teachers in non-tested grades and subjects. If we had the measures and capacity to calculate VAM results, we would not call them "non-tested subjects and grades." However, the authors wisely suggested taking advantage of VAM results where they are available to evaluate other aspects of the educator evaluation system.

One does not even have to adopt the view that subsequent VAM measures are the appropriate criterion for judging educator evaluation systems as advocated by the Brookings report to make use of the authors' suggestions. Yet, we argue that VAM or SGP results can provide information useful for understanding how approaches for incorporating student performance information in non-tested grades and subjects are working. Therefore, we recommend that as part of any pilot and early implementation period the approaches designed for measuring student performance

(growth) in non-tested subjects and grades should be tried out in the tested grades as well. In fact, this fits with our earlier recommendation of using SLOs as a framework for all grades and subjects. This should not be limited to SLOs, but all approaches for calculating "student growth" in non-tested subjects and grades should be included in such an evaluation. If districts are going to design pretest/posttest approaches for documenting student growth in middle school social studies, for example, they should also consider designing a similar system for language arts and mathematics where state tests can be used in the calculation of VAM or SGP. This way, the district (or state) can evaluate the relationship between the "non-tested" approach and the VAM/SGP results. Obviously, these relationships cannot be calculated in truly non-tested subjects and grades, so it makes sense to use the VAM/SGP results that we are able to calculate to help us better understand and evaluate the how the measures used in non-tested grades might be working. Again, Glazerman, et al. (2011) offer one set of criteria and approaches for considering such evaluations. We encourage districts, with state leadership, to consider the general framework suggested by the Brookings report, but to adopt criteria that make the most sense for the specific context.

Conclusions

> *The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor* (Campbell, 1976[11]).

In this paper, we outlined some of the issues and challenges associated with incorporating student achievement results into educator evaluations. We argued that the theory of action is

[11] Campbell, Donald T., Assessing the Impact of Planned Social Change The Public Affairs Center, Dartmouth College, Hanover New Hampshire, USA. December, 1976.

necessary for designing a coherent and valid system and can be used to reveal irreconcilable conflicts between policies and reality. The theory of action is also designed to reveal both the intended consequences of the policies and programs as well as the negative unintended consequences.

One of the major threats to a system that uses assessment results from non-tested grades and subjects is the potential corruptibility of measures used for high stakes purposes that are under the control of those being held accountable. This is a serious concern that threatens the integrity of the entire system and swamps any measurement and analytic concerns. The quote from Donald Campbell above has become known as Campbell's Law for good reason. Recent evidence of cheating on accountability tests in Atlanta, Washington, DC, and several other sites indicates that Campbell's Law is alive and well. It should be noted that most of the recent cheating allegations were on tests for school accountability such the Adequate Yearly Progress (AYP) provision of No Child Left Behind. We can only imagine the pressure that will be felt by teachers when their own jobs are on the line as we move to educator accountability systems. Therefore, it is incumbent on designers of these systems to recognize these threats and try to attend to them in the design. Assuming that all people will act honorably when their jobs are at stake and the system is designed in ways that inadvertently incentivize less than honorable behaviors is turning a blind eye toward a well-documented phenomenon. A theory of action can expose many of the components and mechanisms where teachers might be tempted to act in less than honorable ways. We do not even include teaching to the test in this category, because in most cases teachers think they are doing what is expected of them. On the less honorable side

for example, teachers could subtly (or not so subtly) suggest to their students not to try very hard on the pretest in order to exaggerate the differences between pre and posttest scores.

This is one of the main reasons why we advocated using Student Learning Objectives as an overarching framework for how to incorporate student performance results into educator evaluation systems. It can serve as a means for fostering the types of behaviors that most would like to see, while building the evaluation system. However, even with the use of SLOs, it is not clear that policies requiring that the student growth component count for up to 50% of the effectiveness rating are sustainable, especially for teachers from non-tested grades and subjects. Again, the theory of action is a vehicle for examining the assessment use in a comprehensive manner and we encourage states to consider how the stakes can be balanced with the assessment and analytic quality.

References

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R. J., and Shepard, L.A. (2010, August). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A Preliminary theory of action for summative and formative assessment. *Measurement, Interdisciplinary Research and Perspectives, 8,* 70-91.

Bennett, R. E., Kane, M. & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment.* Paper presented at the 2011 Invitational Research Symposium on Through-Course Summative Assessments.

Braun, H., Chudowsky, N., and Koening, J. (2010). *Getting Value Out of Value Added. Report of a Workshop*. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability. Washington, DC: National Academy Press.

Briggs, D. C. & Domingue, B. (2011). Due diligence in the use of a value added model to evaluate teachers: A review of the analysis underlying the effectiveness rankings of Los Angeles Unified

School District teachers by the Los Angeles Times. National Education Policy Center Working Paper. Retrieved from www.NEPC.org on January 15, 2011.

Bulkley, K. E., Nabors Oláh, L., and Blanc, B. (2010). Introduction to the Special Issue on Benchmarks for Success? Interim Assessment as a Strategy for Educational Improvement. *Peabody Journal of Education: Issues of Leadership, Policy, and Organizations, 85, 2,* 115-124.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., Whitehurst, G. Croft, M. (2011, April). *Passing Muster: Evaluating Teacher Evaluation Systems*. Washington, DC: The Brookings Brown Center Task Group on Teacher Quality. Retrieved on May 22, 2011 from: http://www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx

Goren, P. (2010). Interim Assessments as a Strategy for Improvement: Easier Said Than Done. *Peabody Journal of Education: Issues of Leadership, Policy, and Organizations, 85, 2,* 125-129.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18, 4,* 5-9.

Li, Y., Marion, S., Perie, M. and Gong, B. (2010). An Approach for Evaluating the Technical Quality of Interim Assessments. *Peabody Journal of Education: Issues of Leadership, Policy, and Organizations, 85, 2,* 163-185.

Marion, S. F. & Perie, M. (2009). Validity arguments for alternate assessments. In Schafer, W. and Lissitz, R. (eds.) *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 115-127). Baltimore, MD: Brooks Publishing.

Medina, J. (2010, October). On New York school tests, warning signs ignored. *The New York Times*. Published October 10, 2010. Retrieved on October 11, 2010 from http://www.nytimes.com/2010/10/11/education/11scores.html?ref=education.

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *National Bureau of Economic Research*. Working Paper 14666.

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

Shepard, L. A. (2010). What the Marketplace Has Brought Us: Item-by-Item Teaching With Little Instructional Insight. *Peabody Journal of Education: Issues of Leadership, Policy, and Organizations*, Vol. 85, No. 2: pages 246-257.

Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, D.C.: The Urban Institute.

Appendix A: Summary of tests of assessment evidence available by teaching responsibility[12].

Personnel defined by end of year state summative assessments available	Personnel Type (Examples)	Considerations for Sources of Evidence
Personnel teaching a core subject area where end of year state assessments measuring content taught in their subject area are available in two adjacent grades	Grades 4 -8 (depending on state tested grades) core subject teachers for literacy and math	1. **End-of-year state tests:** The assessments shall form the basis for value-added or large-scale growth model analyses. 2. **Interim Assessments or NRTs:** selected instruments (locally developed or developed by a test vendor) should be deemed as valid and reliable for the purpose of evaluating student growth; ensure data points from prior year be included in evaluating student growth to be consistent with SGP/VAM approach; ensure growth attribution relevant to teacher's contact period with students (e.g., a teacher assigned to a group of students for only one semester). 3. **Student Growth Objective Approach** for evaluating student gains made based on approved instruments and learning goals. Provide training to assist teachers with developing fair and rigorous objectives; tie objectives to state or national standards where available; encourage development of both SMART and stretch goals; ensure SGOs include addressing priority areas requiring improvement identified by both teacher and instructional leader. a. **Student Artifacts:** develop clear criteria defining required elements that need to be included and evaluated; provide training to scorers and teachers evaluating student growth using artifacts; develop fair and clear standards for evaluating growth. 4. **Proportion of growth evaluation attributed to measures evaluating progress towards reaching state, district and school goals:** establish uniform and fair guidelines for determining percentage of total growth evaluation attributed to each selected measure; establish uniform standards for identifying assessments to be used for each measure.

[12] Thanks to Elena Diaz-Bilello for creating the first draft of this table.

	Interventionists/ specialists with shared responsibility with core subject teachers for improving literacy/numeracy skills of students in grades 4-8 (e.g., response to intervention specialists, ELA, special education teachers)	1. **Interim Assessments or NRTs:** selected instruments (locally developed or developed by a test vendor) should be deemed as valid and reliable for the purpose of evaluating student growth; ensure data points from prior year be included in evaluating student growth to be consistent with SGP/VAM approach; ensure growth attribution relevant to teacher's contact period with students (e.g., a teacher assigned to a group of students for only one semester). 2. **Student Growth Objective Approach:** provide training to assist teachers with developing fair and rigorous objectives; tie objectives to state or national standards where available; encourage development of both SMART and stretch goals; ensure SGOs include addressing priority areas requiring improvement identified by both teacher and instructional leader. a. **Student Artifacts:** develop clear criteria defining required elements that need to be included and evaluated; provide training to scorers and teachers evaluating student growth using artifacts; develop fair and clear standards for evaluating growth. 3. **Proportion of growth evaluation attributed to measures evaluating progress towards reaching state, district and school goals:** establish uniform and fair guidelines for determining percentage of total growth evaluation attributed to each selected measure; establish uniform standards for identifying assessments to be used for each measure.

Personnel teaching in a core subject area where an end of year state summative assessment is available to measure content taught in their classrooms, but no prior year state test is available.	Generally science teachers with state tests only in three grades (e.g., 5,8 and 10) and grade 3 teachers with end of year summative state assessments available for their respective grade	1. **Conditional Status Approach:** This is also a VAM approach and the state needs to establish uniform and fair guidelines for determining which predictor scores should be included to calculate growth scores in each content area (reading, math, writing, and science). 2. **Proportion of growth evaluation attributed to shared SGP/VAM:** establish uniform standards for attributing percentage of student growth evaluation to school-wide results in "tested subjects" (e.g., reading and math)**.** 3. **Interim Assessments or NRTs:** selected instruments (locally developed or developed by a test vendor) should be deemed as valid and reliable for the purpose of evaluating student growth; ensure data points from prior year be included in evaluating student growth to be consistent with a SGP/VAM approach; ensure growth attribution relevant to teacher's contact period with students (e.g., a teacher assigned to a group of students for only one semester). 4. **Student Growth Objective Approach:** provide training to assist teachers with developing fair and rigorous objectives; tie objectives to state or national standards where available; encourage development of both SMART and stretch goals; ensure SGOs include addressing priority areas requiring improvement identified by both teacher and instructional leader. a. **Student Artifacts:** develop clear criteria defining required elements that need to be included and evaluated; provide training to scorers and teachers evaluating student growth using artifacts; develop fair and clear standards for evaluating growth. 5. **Proportion of growth evaluation attributed to measures evaluating progress towards reaching state, district and school goals:** establish uniform and fair guidelines for determining percentage of total growth evaluation attributed to each selected measure; establish uniform standards for identifying assessments to be used for each measure.
Personnel teaching in a core subject area where no end of year state summative	Core subject teachers in the sciences (except for the "tested" grades) and social studies. All grades	1. **Proportion of growth evaluation attributed to school-wide measures based on state tested grades and subjects:** establish uniform standards for attributing percentage of student growth evaluation in reading and math (and other subjects, depending on the state system)**.** 2. **Interim Assessments or NRTs:** selected instruments (locally developed or developed by a test vendor) should be deemed as valid and reliable for the purpose

assessments are currently available to measure content taught in their classrooms.	K-2 and most high school teachers.	of evaluating student growth; ensure data points from prior year be included in evaluating student growth to be consistent with a SGP/VAM approach; ensure growth attribution relevant to teacher's contact period with students (e.g., a teacher assigned to a group of students for only one semester). 3. **Student Growth Objective Approach:** provide training to assist teachers with developing fair and rigorous objectives; tie objectives to state or national standards where available; encourage development of both SMART and stretch goals; ensure SGOs include addressing priority areas requiring improvement identified by both teacher and instructional leader. a. **Student Artifacts:** develop clear criteria defining required elements that need to be included and evaluated; provide training to scorers and teachers evaluating student growth using artifacts; develop fair and clear standards for evaluating growth. 4. **Proportion of growth evaluation attributed to measures evaluating progress towards reaching state, district and school goals:** establish uniform and fair guidelines for determining percentage of total growth evaluation attributed to each selected measure; establish uniform standards for identifying assessments to be used for each measure.
	Resource teachers/specialists with instructional responsibility not directly linked to literacy/numeracy skills of students (e.g., music, arts, and P.E. teachers)	1. **Proportion of growth evaluation attributed to school-wide measures based on state tested grades and subjects:** establish uniform standards for attributing percentage of student growth evaluation in reading and math (and other subjects, depending on the state system). 2. **Interim Assessments or NRTs:** selected instruments (locally developed or developed by a test vendor) should be deemed as valid and reliable for the purpose of evaluating student growth; ensure data points from prior year be included in evaluating student growth to be consistent with CGM approach; ensure growth attribution relevant to teacher's contact period with students (e.g., a teacher assigned to a group of students for only one semester). 3. **Student Growth Objective Approach:** provide training to assist teachers with developing fair and rigorous objectives; tie objectives to state or national standards where available; encourage development of both SMART and stretch goals; ensure SGOs include addressing priority areas requiring improvement identified by both teacher and instructional leader.

		4. **Student Artifacts:** develop clear criteria defining required elements that need to be included and evaluated; provide training to scorers and teachers evaluating student growth using artifacts; develop fair and clear standards for evaluating growth. 5. **Proportion of growth evaluation attributed to measures evaluating progress towards reaching state, district and school goals:** establish uniform and fair guidelines for determining percentage of total growth evaluation attributed to each selected measure; establish uniform standards for identifying assessments to be used for each measure.
Personnel who support instructional programs.	Resource teachers/specialists with indirect (non-instructional) responsibility for improving literacy/numeracy skills of students (e.g., social workers, psychologists, and school nurses).	1. **Student Growth Objective Approach:** provide training to assist teachers with developing fair and rigorous objectives; tie objectives to state or national standards where available; encourage development of both SMART and stretch goals; ensure SGOs include addressing priority areas requiring improvement identified by both teacher and instructional leader. 2. **Proportion of growth evaluation attributed to measures evaluating progress towards reaching state, district and school goals:** establish uniform and fair guidelines for determining percentage of total growth evaluation attributed to each selected measure; establish uniform standards for identifying assessments to be used for each measure.