THE COLORADO EDUCATION INITIATIVE

A SURVEY OF APPROACHES USED TO EVALUATE EDUCATORS IN NON-TESTED GRADES AND SUBJECTS

Katie Buckley, Harvard University
Scott Marion, National Center for the Improvement of Educational Assessment
June 2, 2011

There is a growing effort among states and districts to include student tests scores in teacher evaluations. This is due largely to accumulating research demonstrating the importance of teacher quality for improving student achievement (see for example, Hanuschek & Rivkin, 2010; Kane & Staiger, 2008). The primary lever for encouraging states to tie student achievement results to educator evaluations is the Race to the Top (RTTT) fund, announced in July of 2009, which made eligible grant money of nearly 4.4 billion dollars to those states with applications that most closely adhered to reforms that the Obama administration is trying to incentivize. The eleven states along with the District of Columbia that won RTTT grants proposed using student performance as a "significant" factor in teacher evaluations, as well as using teacher evaluations in decisions regarding hiring, firing, tenure and importantly, compensation.

In addition to RTTT grants, the United States Department of Education (USDE) recently committed $1.2 billion over the next five years to the Teacher Incentive Fund (TIF). This year's applicants were awarded grants based, in part, on their plans to create and implement several measures to identify and reward effective teachers, using measures of student growth. Several of the TIF awards have gone to districts proposing to or already implementing the Milken Foundation's Teacher Advancement Program (TAP). TAP includes components that provide multiple career pathways within schools, along with ongoing professional development for teachers, instructionally focused accountability, and performance based compensation. TAP is widespread among districts, including those in Arizona, Arkansas, Colorado, Louisiana, Minnesota, North Carolina, Ohio, Pennsylvania, South Carolina, Tennessee and Texas. As such, many districts implementing TAP have had teacher evaluation systems that include a performance-based pay component in place for several years.

Value-added models (VAM) have become the leading approach for holding teachers accountable for student performance on standardized assessment results.[2] A value-added model measuring "teacher effects" can consist of any model that attributes change in a student's performance relative to other students or relative to some standard to the teacher. There are many different ways to specify a value-added model; however, in general, all VAMs evaluate the change in performance in student test scores over time against a predicted gain score of student achievement, based on school, district or state-wide averages of similar students. If a student's actual gain score is greater than his or her predicted gain score, the difference is positively attributed to their teacher (i.e., the teacher is given a positive value-added score). Conversely, if a student's actual gain score is less than his or her predicted gain score, the difference is negatively attributed to the teacher.[3]

In order to be able to attribute change in performance throughout the year to a student's teacher, a key component for value-added models is test data from at least two points in time, for each student, in the same subject.[4] As a result of the No Child Left Behind Act (NLCB) requirements, most states administer their state test in grades 3-8, in English Language Arts and Mathematics, providing several grades/subjects in which multiple years of data are available. This provides adequate assessment data for approximately 25-35 percent of teachers, leaving 65-

[2] It is important to note that term "value-added model" is a bit misleading since any model can be used to calculate a value-added measure; what makes it a measure of value-added is about the inference of causality from the model rather then the model used.

[3] It's important to note that the general purpose of value-added models is to make causal interpretations regarding teacher effectiveness. However, there are likely important sources of bias if students (and their parents) select certain teachers or teachers select certain types of schools. A primary way to reduce bias in value-added estimates is to randomly assign students to teachers, such that any factors not under a teacher's control that may influence student performance on assessments are factored out. In the absence of random assignment, researchers generally include control variables, such as student demographics and school characteristics, in an attempt to account for observable (but not unobservable) factors that influence student test scores. See Baker et al. (2010) and Braun (2005) for a comprehensive overview of the problems inherent in using non-randomized student test data in teacher evaluations.

[4] VAM models may include several years of student test data to improve the reliability of teacher value-added scores, or include student and school characteristics to include the precision of the estimates.

75 percent of all teachers without adequate information to calculate a value-added score (Goe, 2010). Consequently, states and districts that have proposed using measures of student growth in teacher evaluations (e.g., through RTTT or TIF applications) must determine how to evaluate those teachers in "non-tested" grades and subjects.

In this paper, we define non-tested grades and subjects as those that do not have both a "pre-test" and "post-test" in the same content area that can be incorporated into their teacher evaluation. Note that for non-subject specific grades, such as in elementary school, an assessment that is administered at the end of the prior year generally serves as an adequate pre-test. For example, the mathematics assessment administered to students towards the end of third grade could very well serve as the pre-test for the fourth grade mathematics teachers' evaluations.[5] However, for subject specific grades, such as those in high school, it is often the case that the prior year's exam for a given subject covers a different domain than the current year's exam, and therefore could not be considered a "pre-test" by our definition.[6] For example, the mathematics exam in 9th grade may cover algebra and the mathematics exam in 10th grade may cover geometry; as such, the mathematics exam in 9th grade might not serve as an adequate pre-test for 10th grade mathematics teachers' evaluations.

This paper details the methods states and districts are using to tie student performance to teachers in both tested and non-tested subjects and grades. Rather then providing a broad review of every state and district, we have attempted to delve more deeply into teacher evaluation approaches by focusing only on those states and districts that have *developed* and *promoted* policies to evaluate teachers using student performance. We relied solely on publicly available data, mostly from state and district department of education websites, along with reports that

[5] Note that the same is not true for the fourth grade science teacher, whereby the third grade math test could only serve as a "predictor test", discussed below.
[6] There is also the issue of differential course-taking patterns among students.

addressed this same topic, in order to conduct our search suitable states and districts and gather evidence for this paper.[7] Table 1A in Appendix A provides details of the methods and measures states and districts are using in their teacher evaluation plans for tested and non-tested grades and subjects.

Sample

To identify the states and districts to include in our survey, we searched state and district websites for evidence of the development and/or implementation of teacher evaluation systems. Two main criteria were used to guide our search: (1) States and district teacher evaluation systems needed to include student performance; (2) Documentation on these systems needed to provide sufficient detail of the student performance component and be accessible on state and district websites. In total, we identified 15 states and districts with sufficient information on their teacher evaluation plans, including Massachusetts, Rhode Island, New York, Maryland, Georgia, North Carolina, Tennessee, Delaware (eight of the 12 RTTT winners), additional states of Colorado and South Carolina, and the districts of Washington DC (also a RTTT winner), New York, NY, Hillsborough, FL, Charlotte-Mecklenburg, NC and Denver, CO.

While some states and districts have been tying student achievement to teacher evaluations for many years (e.g., Denver, Colorado and Hillsborough, Florida), most states and districts are still in the planning phase or are updating their systems to explicitly tie student performance to teacher compensation. Therefore, the approaches listed in this paper may very well change. Nonetheless, what follows is an attempt to synthesize the measures and analytic methods that these states and districts are proposing for evaluating teachers in both tested and non-tested grades and subjects. While our survey of states and districts is by no means

[7] Three additional resources detailing what states and districts are doing to tie student performance to teachers in non-tested grades and subjects include Goe and Holdheide, 2010; Steele, Hamilton, and Stecher, 2010; Watson, Kraemer, and Thorn, 2009.

exhaustive of all sites incorporating student performance into teacher evaluation systems, the sites chosen provide a fairly comprehensive list of student performance measures and methods being considered within the United States.

Existing Policy Requirements

According to the Race to the Top requirements, student achievement must be a "significant" part of teacher evaluation systems and is defined separately for "tested" grades and subjects and "non-tested" grades and subjects:

> (a) For tested grades and subjects: (1) a student's score on the State assessments under the ESEA; and, as appropriate, (2) other measures of student learning, such as those described in paragraph (b) of this definition, provided they are rigorous and comparable across classrooms.
>
> (b) For non-tested grades and subjects: alternative measures of student learning and performance such as student scores on pre-tests and end-of-course tests; student performance on English language proficiency assessments; and other measures of student achievement that are rigorous and comparable across classrooms (U. S. Department of Education, 2010)

The RTTT states surveyed typically define "tested" grades and subjects as those that administer a state assessment in that grade/subject and have assessment data from the prior grade that can provide reasonable pre-test scores. For example, Washington DC's IMPACT guidelines state, "The only teachers in DCPS for whom we have both 'before' and 'after' DC CAS [Comprehensive Assessment System] data are those who teach English or math in grades 4-8. Even though we administer the DC CAS in the third and tenth grades, we cannot calculate value added data for teachers of these grades. This is because we have no 'before' data for their students, as we do not test at the end of second grade or at the end of ninth grade." (District of Columbia Public Schools, 2009, p.6).

A common theme across most of the states and districts surveyed is that they are still very much in the planning phase with their teacher evaluation systems, and have instituted internal

task forces or are working with external organization in order to implement their system. Several RTTT states plan to first implement a pilot system or plan to roll out implementation of the full system gradually, such as Rhode Island, Georgia and Colorado. These states are currently in the process of obtaining stakeholder feedback on the system and evaluating the reliability of the data. Most state and district plans tying student performance to teacher evaluations will not go into effect until the 2012-13 school year or later; even then, it not necessarily the case that high-stakes decisions based on the data will be made for all teachers. Washington DC, however, is moving much faster than many other states in rolling out their teacher evaluation system, IMPACT. This system was first introduced in 2009, and currently has one full year of data.[8] Decisions based on compensation will be made starting in the 2011-12 school year, after two years of data have been collected (The District of Columbia Public Schools, 2010).

Other states and districts have had teacher evaluation systems for years; these systems are generally being updated to more explicitly and reliably tie student performance to teacher compensation. For example, Delaware's Delaware Performance Appraisal System (DPAS) II, which was piloted in 2006, created widespread concern among teachers about the way that student performance was incorporated into the evaluations (Beers, 2006). The state is currently in the process of revamping their system, and is still deciding how to hold teachers accountable for student performance in tested and non-tested grades and subjects (Delaware Department of Education, 2010). Hillsborough, Florida's system, known as Empowering Effective Teachers, was established in 2005 as part of state law requiring districts to award higher performing teachers with bonuses. Hillsborough's system led to the creation of hundreds of new end of course assessments, which have since generated some concern regarding their reliability and

[8] In April of 2010, the Union agreed to allow a pay for performance component using IMPACT data, which led to the supplemental system IMPACTplus.

accuracy (Max, 2007). Hillsborough has been given a grant by the Bill & Melinda Gates Foundation to update its system and plans to construct additional states tests for the 2011-12 school year (Steele, Hamilton and Stecher, 2010).

All RTTT states have committed to incorporating student performance as a "significant" portion of their teacher evaluation systems; this has been operationalized as a weight of 20% (e.g., DE) to 51% (e.g., RI) for teachers in tested grades. Within the student performance component of teacher evaluations, states vary in the weight they assign specifically to value added scores based on student test scores from state assessments, as opposed to other measures of student performance, such as reducing student achievement gaps. For example, in New York, of the 40% of the teacher evaluation system that is based on student performance, 62.5% of the 40% (or 25% of the total) will be based on value added scores from student growth on the state assessment, and 37.5% of the 40% (or 15% of the total) will be based on other measures of student performance, using non-value-added scores. Notably, in NY, the percentage based on value-added scores will not increase to 25% until the year of full implementation, 2013-14; prior to this year, value-added scores are weighted at 20%. Other states, like Colorado and Rhode Island, are leaving the weighting of value-added measures flexible until the system is fully developed.

The extent to which the weighting of student performance is the same between tested and non-tested grades/subjects varies from state to state (and district to district). On one hand, Colorado, Maryland Rhode Island and New York plan to weight student performance by the same amount for teachers in both tested and non-tested grades and subjects.[9] On the other hand,

[9] In some states, however, the overall weight on student test scores is split between value-added scores based on the state assessment and measures of growth based on locally-selected measures. For example, in NY, 25% of the teacher evaluation will be based on value-added scores, and the other 15% will be based on locally selected measures.

Georgia and Washington DC plan to have different weighting for student achievement in tested versus non-tested subjects/grades. In Georgia, teachers in tested grades and subjects will have 30% of their evaluation based on observations, 50% based on value-added scores, 10% based on student achievement gap reduction, and 10% based on other quantitative measures. However, teachers in non-tested grades and subjects will have 60% of their evaluations based on observations, and 40% based on other quantitative measures. Since Georgia has clearly stated in their RTTT application that they will not create new tests or implement additional tests in non-tested grades and subjects, "other quantitative measures" appears to include student surveys, parent surveys, and principal/school-focused surveys (Georgia Department of Education, 2010, p. 99-100).[10]

Districts that are implementing teacher evaluation plans through TIF grants, as part of TAP, or through organizations like Battelle for Kids, have more flexibility as compared to the RTTT states in determining how they incorporate and weight student test scores for non-tested grades and subjects in their teacher evaluation systems. Generally, in TAP schools, teachers in tested grades and subjects have 20% of their evaluation based on student test scores, 30% on school-wide measures of performance, and the remaining 50% on teacher observations. For teachers in "non-tested" grades and subjects, 50% of teacher compensation is based on school wide measures and the remaining 50% is based on teacher observations. However, in some TAP districts, teachers in non-tested areas are allowed to choose whether they want the school-wide measure for math or ELA used in their evaluations, depending on what skills they believe they emphasize more in their classrooms (Watson, Kraemer, and Thorn, 2009). Similarly, teachers co-teaching a class in Battelle for Kids schools are allowed to indicate the percentage of time they

[10] Note that while Georgia will not create new assessments, the state assessment system tests students in each grade from kindergarten through high school.

provide instruction directly to students and therefore split the attribution of the value-added scores by the percentage they deem appropriate (e.g., 40%/60%) (Race to the Top Technical Assistance Network, 2010).

Finally, a commonality across all states and districts surveyed is that they plan to use multiple measures to measure student performance and create value-added scores. These measures and approaches are discussed at length below.

Methods of Incorporating Student Performance in Teacher Evaluations in "Tested" Grades and Subjects

Before discussing the approaches used for the non-tested subjects and grades, we first discuss current plans and existing implementations in the tested subjects and grades. Our intent is to provide some context for the later discussion, and to illustrate the various aspects of a very complex system.

The states surveyed for this report are generally considering one of two models to measure student growth in tested grades and subjects. The first is known as the Educational Value-Added Assessment System, or EVAAS, and the second is known as Student Growth Percentiles (SGPs). EVASS was developed by William Sanders and was first pioneered in Tennessee in 1993: for this reason it is sometimes referred to as the Tennessee Value-Added Assessment System (TVAAS). States including Tennessee, Delaware, North Carolina, and Ohio appear to be using EVAAS to calculate value-added scores for teachers in tested grades and subjects. Teachers' value-added scores are calculated as the difference between the average gain on test scores a teacher's students made from the prior year to the current year, and the average gain within the district (Braun, 2005). This model requires pre- and post-test data, with a scale

that is amendable to gain scores, and if available, up to five years of prior student achievement may be incorporated into the model as well. [11]

Like all value-added models, EVASS does not allow for a causal estimate of teachers' effect on students, since there are many factors that may explain student growth (or lack thereof) that cannot be statistically controlled.[12] In the absence of randomly assigning students to teachers, causal inference based on any value-added model is limited. A criticism of EVAAS is the lack of transparency; the exact model is not well publicized and it can be difficult to explain to teachers and parents how scores were derived (Kupermintz, 2003). Since the software is operated by an external organization, SAS, there are considerable costs associated with having the analyses conducted for each student and teacher.[13]

In recent months, a greater number of states have or are starting to employ student growth percentiles to describe student growth and use these growth descriptions in student and school accountability systems. Student growth percentiles (SGPs) were developed by Damien Betebenner (2008) at the National Center for the Improvement of Educational Assessments (The Center for Assessment or NCIEA), in order to provide a normative measure of student growth. SGPs involve ranking the current change in a student's achievement based upon the current distribution associated with prior achievement scores. Upwards of 20 states have been considering using student growth percentiles in their state and federal accountability provisions,

[11] When prior student achievement is available for a given teacher, the model is known as a "layered model" whereby the teacher effect is a function of student growth in the current year and student growth attributed to that teacher from prior years, and adjusted for student learning for the current year's students attributed to other teachers. Notably, student characteristics are not included in the model.

[12] The gold standard to determine causality is to randomly assign students to teachers, thereby removing an effect of unobserved variables. However, in education, it is rarely the case that students can be randomly assigned to treatment (or in this case, to teachers).

[13] Another measure, used by the Dallas school system, is known as the Dallas Value-Added Accountability System (DVAAS) and is considered in alternative to EVAAS. Unlike EVAAS, it does adjust teacher effectiveness estimates for student background characteristics and school-level factors and it does not incorporate multiple years of prior student performance.

with Colorado, Rhode Island, Massachusetts, New York, Kentucky, Washington, Indiana, and Georgia intending to use this method to provide a measure of student growth for teacher evaluations.

Since this method was not initially designed to hold teachers accountable for student learning, states have not yet determined exactly what criteria they will use to assign teachers a positive value-added score or a negative value added score based on SPG results. However, a teacher's "value-added" will likely be based on the median SGP for the class, where a SGP above 50 indicates greater than expected growth and a median SGP below 50 indicates less than expected growth (Betebenner, 2007). Furthermore, it is likely that states and districts will divide the distribution into more than two categories to obtain more distinct measures of teacher performance.

SGPs are a measure of each student's ranking of change in scores among students with the same academic history, but they do not provide a measure of the amount that a student actually grows from year to year. This can be seen as both a benefit and a limitation. The benefit of student growth percentiles is that the measure does not depend on knowing the magnitude of the change; rather it is based on the relative standing of peers and is therefore generally agnostic to the underlying measurement scale (i.e., it does not require a vertical scale). Moreover, it can allow one to extrapolate the likelihood that the student will meet proficiency in future years. However, there is the potential for SGPs to be misleading; it is possible for there to be negative absolute student growth but a positive SGP, as long as the student's negative growth (i.e. decline) is less than that of other students with the same score history.

There are many other types of value-added models that are under consideration. For example, TAP and Battelle of Kids work directly with schools to help them implement a unique

value-added model based upon their needs. According to The Center for Educator Compensation Reform, "each TAP school works with a value-added "vendor"—an independent consultant, an organization such as SAS EVASS for K–12 (SAS Institute, n.d.), or an internal researcher—who has developed value-added methodologies for the TAP schools or districts with which it works." (Lasagna, 2010). Additionally, the Wisconsin Value-Added Research Center (VARC) has worked with several districts, including New York City, to develop a unique value-added model tailored to the district's setting. Currently, many of the states and districts surveyed provide only a broad description of their value-added model, and have not specified the technical details.[14]

Some of the states and districts surveyed plan to incorporate a school-wide measure of student growth, using a value-added model, into evaluations for all teachers. This is true of Washington DC as well as TAP schools; in TAP schools, up to 30% of a teacher's evaluation in tested grades and subjects is based on school-wide measures of student performance. TAP's use of school-wide measures is based on the belief that "school-wide performance evaluations encourage teachers by creating conditions that focus teacher efforts on professional collaboration, student performance and alignment of school resources" (TAP, 2010).

Finally, many states have proposed using additional quantitative measures based on methods other than value-added models in their teacher evaluations. For example, Georgia will base 10% of teacher evaluations in tested grades and subjects on the reduction of the student achievement gap, which the state defines as "the difference in achievement between any student subgroup ($n \geq 15$) in a given teacher's classroom (or overall roster of that teacher's students) and the highest performing subgroup in the State (based on aggregated performance, by student

[14] Decisions include whether to specify teacher effects as fixed or random, whether to specify teacher effects as cumulative, whether to include student-level, classroom-level and/or school-level covariates, and whether the model will be specified as a gain score model, a covariate adjustment model, or a multivariate model (see Briggs and Domingue, 2011, for a discussion of how sensitive teacher value-added scores are to various decisions regarding the model).

subgroup, at the State level)." (Georgia Department of Education, 2010, p. 99). Likewise, North Carolina has proposed incorporating Lexile scores linked to their state assessment system into teacher evaluations for tested grades and subjects and Colorado is considering mandating an additional measure of growth common to those teaching in the same content area in the same school. In these states, as well as others, the technical details of how these quantitative measures will be calculated and evaluated is unclear, and are likely still being determined.

Measurement Tools for Incorporating Student Performance in Teacher Evaluations in "Non-Tested" Subjects and Grades

Most states and districts are considering a variety of assessment types to provide measures of student performance in non-tested grades and subjects. We've grouped these measures into four categories:

1. Externally created *norm-referenced tests* (NRT), such the Stanford-10 or Terra-Nova, and including standardized exams created for special populations, such as ACCESS for ELL students;
2. Externally created *interim assessments* such as Dynamic Indicators of Basic Early Literacy Skills (DIBELS) or Measures of Academic Progress[15] (MAP);
3. National, state or district administered *end-of-course exams* that are standardized, such as the Advanced Placement (AP) exam or the New York Regents assessment;
4. The use of *school- or teacher-developed measures*, including such tools as locally-created end-of-course tests, common performance tasks or other curriculum-embedded assessments, and student portfolios.

[15] In general, there is not necessarily a clear delineation between norm-referenced tests and interim assessments given that many assessment companies offer a menu of tests throughout the year and at the end of the year, but interim assessments are always designed to be administered at least twice and more generally, at least three times a year.

Norm-Referenced Assessments

Externally created norm-referenced assessments are being considered by several of the states surveyed, including Colorado, Delaware, New York, and Rhode Island. These include NRTs like Stanford-10 and Terra-Nova, that provides a measure of student achievement in comparison to national norms. Delaware and Rhode Island are also considering the use of externally created exams for special populations. The primary exam of this type is ACCESS, which is an annual assessment designed to measure student progress in achieving English language proficiency, ACCESS can be administered in grades k-12, which makes it a useful exam to use in measuring value-added for teachers whose students are English Language Learners (ELLs), given the ability to extract pre- and post-test data.

There are two main advantages of summative exams; first the results are comparable across classrooms and schools since the assessment, scoring and administration conditions are standardized, and second, gain scores can be calculated since companies that produce these assessments generally create an examination for each grade. However, administering these exams for every grade and subject for which they are available can be costly. Recognizing this cost factor, Georgia stated in their RTTT application that they will not administer new summative assessments in non-core areas "because such tests must be developed across multiple courses and subject areas, they are not cost-effective." (Georgia Department of Education, 2010, p. 98). Moreover, aligning new assessments to the current system can be a huge challenge if states do not have a set of defined content standards or curriculum.

Interim Assessments

All of the states that are considering NRTs are also considering interim assessments, including DIBELS and MAP. Interim assessments, as defined by Perie, Marion and Gong (2009)

are “assessments administered during instruction to evaluate students’ knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level” (p. 6). Externally created interim assessments are typically constructed to align with state standards and therefore theoretically align with state administered summative assessments.

Similar strengths and limitations exist with interim assessments as with NRTs. Additionally, there is some concern regarding the technical quality of these exams, particularly regarding how well these assessments align with the full range and especially depth of most the state standards (Perie, et al., 2009). Furthermore, these exams are often created with the intention of informing instruction as opposed to a tool to hold educators accountable, and since validity is tied to specific purposes and uses, the validity of such exams may therefore be called into question.

End of Course Assessments

A majority of state and districts surveyed plan to use end of course assessments for some non-tested subjects and grades. States like Rhode Island, Tennessee and Delaware have included AP exams as one measure of a menu of potential measures for non-tested grades and subjects. However, Delaware cautions that AP exams will only be included “if a valid pre-test can be developed” (Delaware Department of Education, 2010). Colorado, New York, Florida, and the district of Hillsborough, Florida plan to use to state- or district-developed end-of-course assessments. For example, Hillsborough Florida has a created over 500 exams for 429 different courses not tested by the state assessment, including foreign languages, art, music, career/technical education, and physical education.

A benefit of end-of-course examinations is that it allows for comparability across classrooms and schools within a district or state. A potential limitation of end of course examinations is that they may have been created for other purposes and not validated for high stakes accountability uses in a growth context. More pointedly, it is very difficult to create assessments with the technical quality sufficient to support high stakes purposes such as educator evaluation. It is unlikely that many districts possess the capacity necessary to do so without expending significant resources for external expertise.

Locally-Created Assessments

A majority of states are also considering school- and teacher-administered measures to assess teacher effectiveness in non-tested graded and subjects. These measures may vary from homework, to reports, to portfolios. For example, Colorado may use measures of "student artifacts" which they define as "classroom-related materials generated by students" including "homework, project-based reports or products, and pictures of student work" (Diaz-Bilello & Marion, 2011). Likewise, NY is considering "other types of locally selected measures, such as writing portfolios, science experiments, and other performance-based assessments." (New York Department of Education, 2010, p. 174).

An advantage of school- and teacher-administered measures is that they are relatively inexpensive to create since many schools and districts already have requirements regarding student work (and teachers already administer a variety of assignments to students). Furthermore, by including teachers in the development of the measures, some of negative stigma surrounding teacher evaluations tied to student test scores may be ameliorated. An obvious disadvantage of this approach is the uncertain technical quality and potential lack of comparability of the results from measures across classrooms.

Analytic Approaches for Incorporating Student Performance in Teacher Evaluations in Non-Tested Subjects and Grades

States are employing different analytic methods to evaluate teacher effectiveness in non-tested subjects and grades using the measures described above. These methods include:

1. *Value-added models* using a pre and post test score from summative assessments in the same subject[16];
2. *Conditional status models* that rely on a covariate (e.g., a "predictor test") from a different content area than the "post-test" or status test;
3. Attributing *school-wide growth* on a state summative assessment to individual teachers;
4. Employing *student learning objectives* (SLOs-also known as student growth objectives) based on teacher or district established goals that are evaluated using district and classroom-based measures.

Value-Added Models

There are many different methods for attributing student test scores to teachers. Value added models are the most popular, since they are geared towards isolating the contribution of individual teachers to student growth. Since a key piece of such models is the need for pre- and post-test data for each student in order to calculate the gain in student learning over time, a potential solution for non-tested grades and subjects is to create or implement new tests to serve as pre and post-tests.

If the issue were that we simply did not have technically adequate tests for these non-tested subjects and grades and some entity was willing to spend significant sums of money, there

[16] Note that there is variability in exactly what a pre-test means. As previously mentioned, we define it as a test that covers the same domain as the post-test, is given in the beginning of the year or the end of the prior year, and is part of the same assessment system as the post-test.

is little doubt that a testing enterprise could be started to provide external tests in subjects such as science and social studies. Those tests would likely provide data allowing the calculation of growth or value-added quantities. However, there are still many concerns with using summative assessments in current non-tested grades and subjects. First and foremost, state and district resources are simply not available to support such an endeavor. Moreover, there are many challenges posed by lack of agreed upon content standards and varied course-taking patterns (i.e., students often do not follow the same course-taking sequences in most courses other than ELA and math and after 8th grade and it is not even clear that students do so in math, particularly across districts).

Conditional Growth Models

When pre-test data in the same subject do not exist, some states are exploring the approach that we at the Center for Assessment have termed the conditional status approach. States with only end of the year assessments for certain grades, as is the case for many grades that administer AP exams and current NCLB science assessments, are considering this approach. This model uses students' earlier scores in another subject to statistically adjust for current performance on a summative assessment. For example, if no adequate pre-test exists for the 8th grade science test, states may include scores from a prior math or reading test (or both) as "predictor" variables in a regression-based model. Several states, such as New York and Colorado, are considering including this approach as part of their repertoire for teachers in non-tested grades and subjects.

An obvious and worrisome disadvantage of this method is that is such predictor tests do not allow for the measurement of growth in any way, and can only control for prior general performance. While correlations of student test scores among different subjects are often quite

high (e.g.,0.4-0.7), it is still possible that students may be particularly strong in one subject (e.g., geometry) and weak in another (e.g., algebra I), which could increase the error of the estimate of teacher value-added scores and lead to misattributing a decline (or improvement) in performance to the current teacher. Further, it is important that teachers feel like they have control over student outcomes for which they will be held accountable and it is not clear that teachers will feel this level of control if the predictor test is from a different content area than the content area for which they are being held accountable.

School-Wide Growth Models

In the absence of strong pre-and post-test data on newly implemented tests, some states and districts are attributing school-wide gains from the state assessment to individual teachers. Tennessee and Maryland are two such states using this approach. Maryland has stated in its RTTT application, "in any grade or subject for which appropriate assessments for calculating individual student-learning growth are not found to be available, MSDE will aggregate student growth gains — from a baseline to at least one other point in time— for the entire school in mathematics, reading, and science (as measured by MSA for elementary and middle schools) and in algebra, biology, English, and government (as measured by the end-of-course High School Assessments for high schools)." (Maryland DOE, 2010). This approach is also being considered by North Carolina, where evaluations for art and music teachers may include the results from students' math and ELA performance on the state test worth approximately 35% of the evaluation (Zelinski, 2010).

There are two slight variations of the school-wide measure approach occurring in districts. Washington DC plans to base 5% of *all* teacher evaluations- those in tested and non-tested grades and subjects-including special education teachers and non-instructional staff, on

school-wide measures of student performance. And, in TAP schools in South Carolina, school-wide achievement growth can account for up to 50 percent of a teacher's evaluation, however, non-tested teachers are given the choice of whether their rewards will be based on math or reading gains depending on whether they believe they emphasize more math or reading skills in their classrooms (Watson, Kraemer and Thorn, 2009; Chait, 2007).

An advantage of using school-wide measures of student gains for individual teachers is that it has the potential to increase school wide effort towards meeting student achievement goals. However, an obvious disadvantage is that it does not provide a direct measure of the effectiveness of individual teachers. Additionally, it may be considered unfair since teachers in certain subjects and grades may have very limited opportunity to influence school-wide math and ELA scores (RTT TA Network, 2010).

Student Learning Objectives

Finally, many states are considering the use of student learning objectives (SLOs) for grades and subjects where implementing a standardized assessment is infeasible. 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). 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.

Notably, SLOs can be used with externally- or locally-created assessments, including teacher–developed measures. Denver, Colorado and the district of Charlotte-Mecklenburg, North

Carolina are at the forefront of this approach. According to North Carolina's RTTT application, "Through the SLO process, teachers and administrators work together to identify specific Standard Course of Study-related areas of focus for each class, and LEA central office staff audit the plans and their implementation to ensure that they are appropriate and are implemented with fidelity. Progress toward meeting SLOs is measured using standardized tests or school- or district-developed tests." (North Carolina Department of Education, 2010, p. 136). &&&

An advantage of this approach is that it is highly flexible: it can be used across all grades and subjects, with existing measures of performance or adapted to new assessment systems as they are developed. Furthermore, because SLOs are often tied directly to regular practices of teachers' work, it is clear to teachers what must be done in order to meet a given performance target, thereby increasing the credibility of the target and potentially creating greater teacher buy-in of the teacher evaluation system. However, this approach is only as good as the quality of the goals set for each student and will therefore require significant professional development in order to be able to create the learning objectives and ensure that the performance goals set are attainable yet rigorous.

Discussion

States and districts are using several approaches to incorporate student performance into teacher evaluations. In tested grades and subjects, the two main approaches are that of Sander's EVASS model and Betebenner's SGP model. While EVASS is widely used, it is less transparent and more costly to stakeholders. SGPs have been increasing in popularity among states, however they do not provide an actual measure of student growth, and therefore the model's extension to a measure of teacher effectiveness is less straightforward. Alternative models have been proposed by organizations such as TAP, Battelle of Kids, and VARC, and are being considered by states and districts as well.

In non-tested grades and subjects, state and districts are considering adding new norm-referenced assessments, interim assessments and end-of-course assessments, along with school or teacher-developed measures of student performance administered at the classroom level to measure teacher value-added. State and districts have proposed several approaches for tying the results from these assessments to teachers. These include the traditional value-added approach, along with a conditional status model when no pre-test is available. In the absence of any technically adequate pre or post-test, states and districts are attaching school-wide measures of student performance to teachers. Finally, an increasing number of states and districts are considering the use of student learning objectives, a framework which can incorporate a variety of measurement tools and approaches.

As this paper demonstrates, incorporating student performance into teacher evaluations is a complex process, and issues of reliability, validity and fairness can and will certainly arise. Furthermore, these systems will likely have long-term consequences for the composition of the teaching force, a factor that will affect students, particularly those in harder to teach schools. It is therefore critical that states and districts ensure that their systems are developed with thought towards continuous evaluation and improvement. In the other papers in this series, we discuss issues surrounding the measurement tools and approaches in greater detail, and provide several recommendations for states and districts as they move forward with incorporating student performance into their evaluations for teachers.

References

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). *Problems with the use of student test scores to evaluate teachers* [Briefing Paper No. 278]. Washington, DC: Economic Policy Institute.

Beers, D. E. (2006). Delaware Performance Appraisal System Second Edition (DPAS II) Pilot: Year One Report. Oak Park, IL: Progress Education Corporation.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), The future of test-based educational accountability (pp. 155–170). New York: Taylor & Francis

Betebenner, D. (2009). *Growth, Standards and Accountability*. Dover, NH: The National Center on the Improvement of Educational Assessment.

Betebenner, D. (2007). *A Primer on Student Growth Percentiles*. Dover, NH: The National Center on the Improvement of Educational Assessment.

Braun, H. (2005). *Using Student Progress to Evaluate Teachers : A Primer on Value-Added Models*. Princeton, NJ: Educational Testing Service.

Briggs, D. & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/due-diligence.

Chait, R. (2007, November). *Current state policies that reform teacher pay: An examination of pay-for-performance programs in eight states.* Washington, DC: Center for American Progress. http://www.americanprogress.org/issues/2007/11/pdf/teacher_pay.pdf

Delaware Department of Education. (2010). *Matrix of appropriate measures of student growth for DPAS II Component V*. Retrieved January 11, 2011, from http://www.doe.k12.de.us/csa/dpasii/student_growth/files/Matrix_MeasDPASIIComp_V_June_Sum.doc

Diaz-Bilello, E. and Marion, S. (2011). Required and supplementary assessments and measures by personnel type**.** Working document of the State Council for Educator Effectiveness. Colorado Department of Education.

District of Columbia Public Schools. (2009). *IMPACT guidebooks*. Washington, DC: Author. Retrieved January 19, 2011, from http://www.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+(Performance+Assessment)/IMPACT+Guidebooks

District of Columbia Public Schools (2010). Impact PLUS: The District of Columbia Public

Schools Performance-Based Compensation System 2010–2011. Retrieved February 1, 2011, from http://www.dc.gov/DCPS/Files/downloads/TEACHING%20&%20LEARNING/IMPACT/IMPACTplus/DCPS-IMPACTplus-guidebook-Sept-2010.pdf

Georgia Department of Education (2010). Georgia's Race to the Top Application. Retrieved December 28, 2010, from http://www2.ed.gov/programs/racetothetop/phase1-applications/georgia.pdf.

Goe, L. (2010). Teacher Evaluation in Transition: Using Evaluation to Improve Teacher Effectiveness. The National Comprehensive Center for Teacher Quality.

Goe, L. (2010b). Student Growth in Non-Tested Subjects and for At-Risk Students**.** The National Comprehensive Center for Teacher Quality. Retrieved January 10th, 2011 from http://www.swcompcenter.org/educator_effectiveness/Student_Growth_in_Non-Tested_Subjects_and_for_At-Risk_Students,_Laura_Goe.ppt

Goe, L. and Holdheide, L. (2010), Measuring Teachers Contribution to Student Learning Growth for The Other 69%. The National Comprehensive Center for Teacher Quality. Retreived January 23, 2011 from http://www.tqsource.org/webcasts/201012Workshop/ measuringTeachersContribution.pdf.

Hanushek, Eric A., and Steven G. Rivkin. (2010). "Generalizations about using value-added measures of teacher quality." *American Economic Review,* 100(2): 267-271.

Kane, T. J., and Staiger, D. O. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Koedel, C. (2007, September). *Teacher quality and educational production in secondary school.* Working Paper 2007-2. Nashville, TN: Vanderbilt University, National Center on Performance Incentives. http://www.performanceincentives.org/data/files/news/PapersNews/Koedel_2007a_Revised.pdf

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287-298.

Lasagna, M. (2010) *TAP: The System for Teacher and Student Advancement.* Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington, D.C.

Max, J. (2007). *The Evolution of Performance Pay in Florida.* Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington, D.C.

National Comprehensive Center for Teacher Quality. (2010). *Guide to teacher evaluation products*. Retrieved February 13, 2011, from http://www3.learningpt.org/tqsource/gep/default.aspx

New York Department of Education. (2010). Race to the Top Application: Phase 2. Retrieved December 27, 2010, from http://usny.nysed.gov/rttt/application/home.html

North Carolina Department of Education. (2010). Race to the Top Application. Retrieved December 29, 2010, from http://www2.ed.gov/programs/racetothetop/phase1-applications/north-carolina.pdf

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) pp. 5-13.

Race to the Top Technical Assistance (TA) Network. (2010). Measuring Student Growth for Teachers in Non-Tested Grades and Subject: A Primer. Washington DC: ICF International.

Rhode Island Department of Education (2010). Race to the Top: Application for Initial Funding. Retrieved December 28, 2010, from http://www.ride.ri.gov/commissioner/RaceToTheTop/docs/RhodeIsland-RTTTapplicationnarrative.pdf.

Steele, J. L., Hamilton, L. S., and Stecher, B. M. (2010). Incorporating Student Performance Measures into Teacher Evaluation Systems. Santa Monica, California: RAND

TAP: A System for Teacher and Student Advancement. (2010). *Performance Based Compensation.* Retrieved February 10th, 2011 from: http://www.talentedteachers.org/policyresearch/policyresearch.taf?page=elements_pbc U.S)

United States Department of Education. (2010) Race to the Top Executive Summary. Retrieved January 23, 2011, from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Watson, J.G., Kraemer, S.B., and Thorn, C.A. (2009). *The Other 69 Percent*. Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington, D.C.

Zelinski, A. (2010, July 9). *Teacher evaluation panel hashes out first set of recommendations*. Retrieved September 20, 2010, fromhttp://www.tnreport.com/2010/07/ teacher-evaluation-panel-hashes-out-first-set-of-recommendations/

Appendix A

Table A1: Description of Teacher Evaluation Plan for Tested and Non-Tested Grades and Subject, by State and District.

	Tested Grades and Subjects*						**Non-Tested Grades and Subjects***		
State/ District	**Year of full implementation ****	**Year of initial implementation (i.e., pilot or prior system)****	**Percent of evaluations based on VAM scores**	**VAM**	**School-Wide VAM?**	**Non-VAM measures of student performance**	**Percent of evalua-tions based on student test scores**	**Test measures**	**Analytic approach**
DE	2011-2012: Will determine student growth scores for all teachers	2006-07: First implemented, but without student performance component	20%	EVAAS	Unclear	Potentially several, including use of student portfolios	20%	NRTs, interim assessments	Unclear
TN	2011-12: Student growth will be used to make decisions regarding teachers	1992: State-wide value-added system was put in place	35% (with an additional 15% based on other measures of performance)	EVAAS	Unclear	Other measures of student achievement such as reading assessments for elementary teachers and college entrance tests, end-of-year subject tests and advance-placement tests for high school teacher	Unclear	Unclear	Measures of school-wide growth, new end of course assessments at the high school level
MA	2013-14	Unclear	Unclear	SGPs	Unclear	Unclear	Unclear	Unclear	Student work samples

	Tested Grades and Subjects*						Non-Tested Grades and Subjects*		
State/ District	**Year of full implementation ** **	**Year of initial implementation (i.e., pilot or prior system)** **	**Percent of evaluations based on VAM scores**	**VAM**	**School-Wide VAM?**	**Non-VAM measures of student performance**	**Percent of evalua-tions based on student test scores**	**Test measures**	**Analytic approach**
NY	2013-14: Following Regents approval for value-added model for all teachers	2011-12: Projected pilot year for tested grades and subjects (4th-8th. Mathematics and ELA)	20%-40% on student growth on state assessment (with remaining percent based on other measures of student performance)	Not clear, RFP just released	Unclear	Yes, locally selected measures	40%	Unclear	Potential pre- and post-tests, end-of-course tests, and performance on English language proficiency assessments,
RI	2012-2013: RI will calculate student growth data for all teachers with rewards/consequences	2010-11: Projected small pilot in several districts	51%: (But unclear what percentage will come from value-added scores vs goal attainment vs school-wide measures)	SGPs	Yes	Yes, locally and district-wide tests	51%: Using a goal attain-ment process along with school wide measures	Locally selected measures including NRTs and district wide tests (e.g., NWEA, AP exams)	Goal attainment process, measures of school/group-wide growth
MD	2012-13	Unclear	50%	Unclear	Unclear	Unclear	50%	Pre-and Post tests already used in schools	Measures of individual growth, measures of school-wide growth

	Tested Grades and Subjects*						Non-Tested Grades and Subjects*		
State/ District	**Year of full implementation ** **	**Year of initial implementation (i.e., pilot or prior system)** **	**Percent of evaluations based on VAM scores**	**VAM**	**School-Wide VAM?**	**Non-VAM measures of student performance**	**Percent of evalua-tions based on student test scores**	**Test measures**	**Analytic approach**
GA	2012-2013	2011-2012: Validate survey tools and field test other quantitative measures	50%	SGPs	Unclear	Reduction of student performance gap (10%)	0% (Georgia will attribute 60% of evalua-tions to observa-tions and 40% to surveys.)	No new tests will be created.	None
NC	2012-13: High stakes decisions will be made as long as educators have at least three years of data	2010-12:Teacher effectiveness initiative study of various measures	Unclear	EVAAS	Unclear	Yes: ABC Growth Measures	Unclear	Locally developed pre- and post tests, ABCs, IEPS and AMOs	VAM, student learning objectives
SC (TAP Schools)	2010-11 : Expansion to 16 districts	2002-3: Pilot in 5 districts and 6 schools	30%	Unclear	Yes: 20%	Unclear	50%: But for school-wide value-added measures	School-wide measure of student perform-ance	Unclear

	Tested Grades and Subjects*						Non-Tested Grades and Subjects*		
State/ District	**Year of full implementation ****	**Year of initial implementation (i.e., pilot or prior system)****	**Percent of evaluations based on VAM scores**	**VAM**	**School-Wide VAM?**	**Non-VAM measures of student performance**	**Percent of evaluations based on student test scores**	**Test measures**	**Analytic approach**
CO	2013-14	2011-12: Projected pilot year	50%	SGPs	Unclear	Measure common to those teaching in the same content area	50%	Depends on type of personnel. Tools include NRT, interim assessments and teacher created tasks	Depends on type of personnel. Methods considered include conditional status models and student learning objectives
DC	2012-13: Non-tested grades and subjects will likely be included at this point	2009: Data began to be collected for teachers in tested grades and subjects, with high stakes decisions occurring in 2010	50%	Covar-iate adjusted model using multiple reg-ression	Yes: 5%	Unclear	10%	Teacher chosen measure	Unclear
New York, NY	2014-15	2011-12	40%	Likely SGP	Unclear	Unclear	Unclear	Unclear	Unclear
Hills-borough, FL	2011-12: Once new end of course assessments have been developed	2005-06: Pilot year	40%	Unclear	Unclear	Unclear	40%	Creation of new end of course assessments	Value added models of student growth

	Tested Grades and Subjects*						Non-Tested Grades and Subjects*		
State/ District	**Year of full implementation ****	**Year of initial implementation (i.e., pilot or prior system)****	**Percent of evaluations based on VAM scores**	**VAM**	**School-Wide VAM?**	**Non-VAM measures of student performance**	**Percent of evalua-tions based on student test scores**	**Test measures**	**Analytic approach**
Charlotte Mecklen-burg, NC	Unclear	Unclear	Unclear	Likely EVAAS	Unclear	Unclear	Unclear	Unclear	Measures of school-wide student growth, Student Learning Objectives
Denver, CO	Unclear	2005-06: First Implemented	50% (but uncertain)	Likely SGP	Yes	Teacher selected assessments	Unclear	Teacher Selected Assess-ments	Measures of individual student growth, measures of school-wide student growth, possibly Student Learning Objectives

*Information gathered from state websites, RTTT applications and articles cited in paper. Data is current as of March, 2011.

**Please note that states and districts typically have different definitions for system implementation. For some it means when they will start collecting data on teachers, for others it means when they will use the data to make high-stakes decisions about teachers in tested and/or non-tested grades and subjects.