STATE PRACTICES RELATED TO THE USE OF STUDENT ACHEIVEMENT MEASURES IN THE EVALUATION OF TEACHERS IN NON-TESTED SUBJECTS AND GRADES[1]

Erika Hall

Douglas Gagnon

Jeri Thompson

M. Christina Schneider

Scott Marion

National Center for the Improvement of Educational Assessment

August 26, 2014

[1] We are grateful to the Bill & Melinda Gates Foundation for supporting the production of this paper. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Foundation.

State Practices Related to the Use of Student Achievement Measures in the Evaluation of Teachers in Non-Tested Subjects and Grades

In recent years, states and districts across the nation have been working to incorporate indicators of student achievement into their existing or newly envisioned teacher evaluation systems. This push was initially due to state efforts to obtain Race to the Top (RTTT) funds, but took hold more recently as states struggled to meet Elementary and Secondary Education Act (ESEA) waiver requirements. To date, 43 states plus Puerto Rico and the District of Columbia have received ESEA waivers, which necessitate —among other things—the inclusion of measures of student growth[2] in teacher evaluations. The increased use of student achievement as an indicator of teacher effectiveness reflects a growing belief that effective teachers should have an observable, measureable impact on student achievement, and rests on a theory of action holding that the inclusion of measures of student achievement in the evaluation process will have a positive impact on student learning.

Three of the greatest challenges faced by states and districts as they work to incorporate measures of student achievement into their teacher evaluation systems include identifying or developing appropriate assessments, analyzing student data, and determining how to interpret and combine results to make valid inferences about a teacher's impact on student learning. While these challenges apply to all teacher evaluations, they are especially problematic when defining procedures for teachers associated with subjects and grades for which high-quality student performance data from state-developed standardized assessments are not readily available. For teachers in these non-tested subjects and grades (NTSG) the lack of such data precludes the use of typical approaches to estimate a teacher's impact on student learning, including value-added modeling (VAM) and student growth percentiles (SGP). Such approaches strive to estimate the unique contribution of a teacher to student learning (i.e., value added)[3]

[2] Although student growth often carries a specific meaning in the assessment field (ie., change in performance along a vertically-scaled domain) within the context of this paper, growth is used to broadly refer to any difference or change in student outcomes used to support teacher evaluation (i.e., change in performance from pre-test to post-test; difference between expected and observed performance, etc.)

[3] In this paper, the phrase value added refers to the contribution of a given teacher to student learning, independent of the means by which that contribution is estimated.

by comparing observed performance to expected student performance after accounting for prior achievement. These and related approaches are referred to as conditional status models within the context of this paper.

Despite the fact that teachers associated with NTSG typically comprise the majority of all teachers (Prince et. al., 2009; Buckley & Marion, 2011) the development of comprehensive evaluation procedures for this population has greatly lagged behind that of other teachers. This is primarily due to difficulties in obtaining "measures of student achievement that are rigorous and comparable across schools within a district," which can stem from a variety of factors unique to NTSG. For example, in some subjects content standards may be non-existent or poorly specified. This is often occurs in areas such as drama and physical education where teachers are given greater flexibility to determine the type and level of skills students should develop within a given grade or over the course of a school year. Even for those subject areas where clear content expectations exist, the specific content and skills taught within a given grade may differ from school to school, making the development of assessment tasks and/or procedures that provide for comparable measures across schools extremely difficult.[4]

Another common issue for NTSG is a lack of state and/or district resources (e.g., monetary and human) to support the selection or development of appropriate, high-quality assessments; professional development of teachers around the identification and collection of appropriate student performance data; and precision in the analysis of obtained results by state/district staff — all of which are necessary to provide for fair and accurate estimates of student achievement. Subject areas that cannot be assessed using paper-pencil or computerized tests are especially problematic as they often require the development of complex performance tasks that can be expensive to develop, time consuming to administer, and difficult to score.

[4]While the need for "comparable" measures is outlined as an ESEA waiver requirement, the way comparability is defined and the role it should play in evaluating the quality and appropriateness of a given measure for teacher evaluation is debatable. From our perspective, if two or more measures provide for valid inferences about a teacher's impact on student learning within a given subject area or course, those measures are comparable for the purposes of supporting educator evaluation – even if they are not comparable from a strictly psychometric point of view.

In 2011, Buckley and Marion summarized the tools and approaches being used, or proposed for use, by select states and districts in evaluating teachers in both tested subjects and grades (TSG) and NTSG. At that time, the number of states and districts with documented or preliminary plans in place was extremely limited, consisting mainly of states and districts that had received RTTT or other external funds (e.g., Teacher Incentive Fund), and little thought had been given to the development of procedures appropriate for NTSG. Now, approximately three years later, most states have a comprehensive plan in place to support teacher evaluation, and these plans include at least some specifications for NTSG. Although such plans differ greatly from state to state (e.g. timelines for implementation, degree of specificity, etc.), it is apparent that states and districts have explored a variety of options with respect to the assessments, tools and data used as evidence of student achievement, and the approaches used to appropriately attribute those measures to individuals or groups of teachers.

Purpose

The purpose of this paper is to summarize current state practices on the collection and use of measures of student achievement for teacher evaluation in NTSG, and to identify key areas of variability across states and districts with respect to:

- the assessments, tools and data used to establish or represent measures of student achievement;
- the approaches used to attribute student achievement measures to educators for the purpose of supporting teacher evaluation;
- the degree of local control afforded to districts in designing, implementing and evaluating teacher evaluation procedures for NTSG; and
- key policy factors that contribute to the manner in which evaluations systems for NTSG are specified, evaluated and implemented.

To address ESEA waiver submissions, states have given a great deal of thought to how evaluation might be conducted for teachers in NTSG; therefore, additional and improved documentation related to planned procedures and goals for teacher evaluation is publically

available for review and analysis. In addition, since RTTT funds were awarded beginning in 2010, many Phase 1 and 2 winners have had teacher evaluation systems in place for multiple years. This has allowed for the review and refinement of existing evaluation procedures as well as documentation of the pros and cons associated with each.[5] Similarly, although USED has relaxed requirements related to when student achievement data must be used to inform decisions about teachers, many states and districts have been piloting components of their teacher evaluation system for several years to meet the initially proposed 2015–2016 deadline.

As a result of these activities, states have a much stronger foundation upon which to base decisions on the fairness, feasibility and appropriateness of evaluation tools and techniques for teachers in NTSG than they did three years ago. Consequently, proposed evaluation procedures are more comprehensive and less tentative than those initially envisioned during the ESEA waiver process or summarized back in 2011. All of these reasons have impressed upon us the need to extend the work initially presented by Buckley and Marion (2011).

Definition of Non-Tested Subject and Grades

Because what constitutes a NTSG often differs between states and districts, for the purposes of this paper a NTSG teacher is one who *does not receive a state-supplied score reflecting student growth or teacher value-added in light of his/her student's performance on a common, standards-aligned assessment in the primary subject area for which they teach*. Given this definition, if a state develops and annually administers a common, standardized, end-of-course (EOC) assessment within a traditionally NTSG (e.g., Biology, Fine Arts, Physical Education, etc.) and the state produces a growth measure that supports teacher evaluation in light of student performance on that test, the subjects and courses associated with those EOC exams would no longer be considered non-tested. For example, over the last several years North Carolina has expanded common exams to include all four core subjects in grades 4–12, which encompass roughly 70 percent of all teachers in the state. Although only 40 percent of these teachers currently receive a VAM-based estimate, these assessments provide for a common measure of student growth that supports teacher evaluation and therefore, given our current definition,

[5] This is also true for a variety of districts across the country that were awarded Teacher Incentive Fund grants.

the subjects and courses associated with these assessments would now be considered "tested." In North Carolina, subject areas that do not have a common assessment (e.g. Arts, World Languages, etc.) are measured through an Analysis of Student Work Process, whereby teachers collect student work artifacts, assess them, and then submit them for "blind review" by another teacher in the state to determine the amount of growth observed. Since the growth estimates used to support teacher evaluation in these subjects are not based upon student performance on standards-aligned assessments, they would still be NTSG for our purposes.

Procedures and Methodology

A multi-phase process was used to establish the data and information outlined in this paper. First, a team of researchers from the Center for Assessment worked to outline a set of questions that would inform our understanding of how states were evaluating teachers in NTSG. After multiple iterations, the set of 17 questions presented in Appendix A became the focus of our analysis. Next, to gather information specific to each state and the District of Columbia, we reviewed state Department of Education (DoE) websites for state laws and statutes related to teacher evaluation, relevant policy documentation, teacher evaluation resources and implementation guidelines, ESEA waiver applications, and any other pertinent documentation. In some cases, documentation outlining procedures specific to teachers in NTSG was non-existent, limited, out-of-date, or described inconsistently across sources. Therefore, to ensure confidence in the accuracy of the information collected, each state's DoE was subsequently contacted and asked to review, validate and/or correct the information.[6] A total of 30 responses were received, in which states extended, corrected or clarified the information collected. While this greatly increases our confidence in the accuracy of the results presented below, it raises concerns about the information collected for the 21 non-responders. As a result, most results are presented in terms of general trends rather than exact counts or percentages.

[6] A sample state summary, which includes the manner in which particular terms and vocabulary were intended to be interpreted is provided in Appendix B.

Early in this process, it became clear that many of the terms and concepts commonly used to describe teacher evaluation practices are defined and operationalized differently across states. For example, in some states student achievement measures (SAM) are conceptualized as measures of student performance resulting from the administration of an assessment or tool aligned to relevant content standards. Other states, however, define SAM much more broadly, including measures of school performance (e.g., grades/scores used for accountability), attendance rates, Advanced Placement (AP) participation, and other factors thought to be indicative of student performance (at a class or school level) in addition to student performance on assessments. This distinction is important, because a state may utilize a variety of SAM to support inferences about teacher effectiveness, but not all provide for estimates of an individual teacher's impact on student learning. Some measures support the attribution of student learning to the teacher, while others are simply indicators of student achievement considered important enough to be included in the evaluation system. In light of such discussions, staff members worked together to operationalize ambiguous terms and concepts to ensure they were interpreted consistently.

Summary of Key Findings

Data resulting from the state summaries are presented in two sections. The first section reflects responses to those questions for which there was little to no variability between expectations for teachers in TSG and NTSG. That is, responses were not specific to the design and implementation of evaluation systems for teachers in NTSG, but common to all teachers in the state. It includes items such as the year in which the system must be fully operational, whether the state allows districts to develop or propose alternate systems of evaluation, the percentage of a teacher's overall evaluation based upon measures of student achievement, and whether the state provides districts with a comprehensive model or a set of guidelines to support teacher evaluation. The second section summarizes key trends across states regarding the procedures and information used to support teacher evaluation for NTSG, and the degree of local control afforded to districts in defining and implementing components of their evaluation system for teachers in this population. While the results summarized below are

based upon information collected through state responses to all 17 questions outlined in Appendix A, for ease of interpretation and presentation, key elements relative to each state are summarized in the table presented in Appendix C.

Although the use of SAM to support teacher evaluation in TSG was not the focus of our research, a couple observations related to the procedures used for TSG and NTSG are worth highlighting. While such variations are not in and of themselves surprising, their implications, especially as they relate to the degree of local control afforded to districts, have not previously been discussed. Therefore, when appropriate, key differences observed in the evaluation of TSG and NTSG are also addressed.

Deadlines for Teacher Evaluation Systems to be Fully Operational

As discussed in the previous section, the design of evaluation systems for teachers in NTSG has lagged behind that of teachers in TSG. Despite this, state deadlines related to operational implementation are generally the same for teachers in NTSG and in TSG. Table 1 offers a breakdown of the dates indicated by states as to when educator evaluation systems should be fully operational (shown also in column 2 in Appendix C).

Table 1. Dates Fully Operational[7] Across States[8]

Fully Operational	RTTT Winners	Non RTTT Winners	Total
2013-2014 or earlier	10	4	14
2014-2015	8	13	21
2015-2016 or later	1	8	9
TBD	0	2	2
NA – No ESEA Waivers	0	5	5
Total	19	32	51

Within the context of this report, "fully operational" does not necessarily mean that all components of the system are in place or all elements are being weighted as intended. For

[7] For the purposes of this study, we consider "fully operational" to mean that a teacher evaluation system/model/guidelines is/are in place for all teachers.

[8] Note that results for DCPS are also included in these analyses.

example, the state of Pennsylvania is fully operational as of 2013–2014, but estimates of student achievement based upon the use of a Student Learning Objective (SLO) process are not required until 2014–2015 for teachers in both TSG and NTSG. Similarly, fully operational does not imply that evaluations will result in consequences for teachers that year, as many states results will not use results to support personnel decisions for one or two years after the system is in place for all teachers. It simply indicates that all teachers in the state must *fully participate* in a state or district approved evaluation system during that year, regardless of how it is currently defined.

The results in Table 1 show that most RTTT states are ahead of non-RTTT states in terms of when systems are planned to be fully operational. For example, nine of the twelve Phase 1 and 2 RTTT winners had fully operational systems in place as of the 2013–2014 school year, and the remaining three are scheduled to be fully operational in 2014–2015. Delaware and Tennessee, the first RTTT winners, have had systems in place for all teachers since 2012–2013 and 2011–2012, respectively. The accelerated timelines reflected by RTTT states is not surprising given they have had greater time and resources to support the design, piloting, and revision of their systems, and most have enacted statutes and/or regulations to ratify RTTT application requirements.

Provision of Locally Developed Alternate Teacher Evaluation Systems

Most states provide districts with a great deal of flexibility in defining their teacher evaluation systems. While the type and degree of control afforded may differ for TSG and NTSG, the decision as to whether alternate systems or deviations from a state-defined model will be considered or allowed is a policy decision that occurs at the state-level. For the purposes of this paper, a *model* was defined as a comprehensive state-developed template or vision for evaluation that could be implemented by a district as an intact system with little need for design or modification. On the other hand, *guidelines* were broadly interpreted as a set of rules, recommendations, examples, or narratives provided by the state to support districts with system design and implementation. The distinction between models and guidelines is often quite murky. Some states consider their guidelines specific enough to be a model, while others

provide a model but give districts the option of developing an alternate system and are therefore are utilized as guidelines. Currently, 80 percent of the states allow for districts to either develop their own evaluation system (in consideration of provided guidelines) or propose modifications to all or some components of the state-defined model. Of these, however, two-thirds require some level of state review or approval prior to district implementation.

Percentage of Total Evaluation Dependent on Measures of Student Achievement

Of the 35 states that identified an explicit percentage for SAM in teacher evaluation, only three indicated different weights for teachers in TSG and NTSG, and in each of the cases the percentage was lower for teachers in NTSG. For example, in Washington, DC Public Schools, 15 percent and 50 percent of a teacher's overall evaluation is proposed to be based upon SAM for NTSG and TSG, respectively. Table 2 summarizes the indicated percentage of a teacher's overall evaluation that is based in some way upon measures of student achievement.

Table 2. Percentage of Evaluation Based Upon Measures of Student Achievement

Designated Percentage	RTTT Winners	Non-RTTT Winners	Total
15-25	4	9	13
30-40	2	3	5
50	10	7	17
Not Specified	3	13	16
Total	19	32	51

There is clearly some variability across states with respect to the emphasis given to SAM in a teacher's overall evaluation. Designated percentages range from 15 to 50 percent, with roughly one-third of states suggesting that results based on SAM account for half of a teacher's overall effectiveness rating. In general, the weight associated with SAM was greater in RTTT states than among non-RTTT winners. Over half of RTTT states suggested that SAM represent 50 percent of a teacher's overall evaluation, compared to approximately 20 percent of non-RTTT winners.

States for which percentages were not specified either left this decision up to districts or had not yet made a decision with respect to weight. In a few states, SAM are not assigned a percentage, as they are considered a "gateway" necessary to achieve an overall effective designation (e.g., North Carolina) or a "trigger" to alter the overall rating if a discrepancy between teacher/student performance is observed. For example, in Arkansas a teacher's overall effectiveness rating is equivalent to his/her Professional Practice rating, unless the median Student Ordinal Assessment Ranking (similar to a SGP) does not reach a state-defined growth threshold. Under these conditions a teacher's overall rating cannot be "Distinguished" regardless of the Professional Practice rating earned. Furthermore, if a teacher does not meet the growth threshold two years in a row, his/her overall effectiveness rating is automatically lowered one level.

A couple points should be kept in mind when interpreting the results in this table. First, the percentage represented does not always reflect the percentage of a teacher's evaluation based solely on estimates of student growth. For example, both Pennsylvania and Georgia indicate that 50 percent of a teacher's overall evaluation is based upon measures of student achievement. However, in Pennsylvania that 50% is comprised of teacher specific measures of student growth (as obtained through an SLO process or other approved procedure) in addition to a state-developed school performance index. While in Georgia the entire 50 percent is based upon measures of student achievement that provide for estimates of student growth.

Second, although four out of five states refer to f SAM as a percentage of a teacher's overall evaluation, what this percentage means, in terms of impact on a teacher's final evaluation, depends on the procedures used to establish the final effectiveness rating. For example, in Louisiana a teacher's overall effectiveness rating is based upon a weighted sum of the contributing components, with a teacher's growth score weighted by 50 percent. In this case, from a purely mathematical stand-point, half of the teacher's final rating is based upon SAM.[9] On the other hand, Hawaii indicates that 50 percent of a teacher's overall rating is based upon

[9] It is important to note that even though the nominal weight assigned to SAMs is 50 percent, the effective weight may be smaller or larger depending on the observed variability of that and other measures included in the model. (See Kolen and Brennan, 2004)

SAM but a two-variable decision matrix (teacher practice and student growth) is used to determine a teacher's overall rating. While student growth does represent 1 of the 2 components (50%) ultimately used to assign a final effectiveness rating to Hawaii teachers, the relative impact of this factor depends on the manner in which the data are combined across the different components of the system.[10] If a compensatory approach is used, such that high performance on one component of the system compensates for low performance on the other, and this occurs in an equivalent manner across components (e.g., a rating of Effective in student growth, compensates for a rating of Moderate in professional practice and vice versa), then the assumption of equal weighting may hold. If, however, a conjunctive approach is applied, such that a certain threshold or level of performance must be obtained in one area (e.g., growth) in order for an overall rating of effective to be obtained (regardless of performance in other areas), discussing the impact of that component relative to the total number of components in the system does not make sense.

From our review and the feedback provided by states, we estimate that 15 states require or recommend the use of a decision matrix to combine system-based measures and determine a final effectiveness rating; 17 recommend an additive procedure; and 18 allow districts to independently determine how components should be combined.

Assessments Tools and Data used to Establish or Provide Evidence of Student Achievement

One of the major challenges facing states and districts related to the evaluation of teachers in NTSG relates to the selection of assessments or identification of data appropriate for use in providing evidence of student achievement. As a result, a broad range of measurement tools and results have been proposed for use within and across states. Some of the most common are represented in Table 3 below.

[10] See Diaz-Bilello,E., & Hall, E. (2014) for a discussion on different techniques of combining results across components of an Educator Evaluation (EE) system. **[Note: EE system should be defined—ab]**

Table 3. Sources of Student Achievement Data for NTSG

Source of Student Achievement Data	Examples
State developed and administered large scale standardized assessments, (i.e., those associated with TSG)	Pennsylvania State System of Assessments (PSSA); Hawaii State Assessment (HAS)
State or district administered End-of-Course (EOC) assessments that are standardized	Advanced Placement (AP) exams, New York Regents Assessments; Tennessee's US History EOC Assessment
District, school, or teacher developed assessments, including locally developed EOC tests, performance-based assessments, student portfolios	Pre and Post-test assessments developed by districts in Georgia, EOC assessments developed by Hillsborough County, Florida.
Nationally recognized, large-scale norm-referenced assessments	ACT or SAT suite of assessments, Terra Nova, SAT 10
Vendor developed off-the-shelf interim, benchmark or diagnostic assessments	Dynamic Indicators of Basic Early Literacy Skills (DIBELS), Developmental Reading Assessment (DRA), Measures of Academic Progress (NWEA MAP)
School-Based Performance Measures	School accountability measures; Graduation/attendance/drop-out rates etc.; Student or parent survey results; AP/ACT/SAT/IB participation; Passing rates on state assessments; Achievement gap reduction

Of the sources listed above, those in the first five rows provide for measures of student achievement that also allow for inferences about an individual teacher's impact on student learning (i.e., value added). However, roughly one in five states also allow or require the use of school-based performance measures such as those presented in row 6, to support teacher evaluation. For example, while Arizona specifies that 33–50 percent of a NTSG teacher's evaluation should be based on measures of student achievement, only 20 percent of the total evaluation outcome is required to consist of an academic growth calculation between two points in time. The remainder may be represented by any of a variety of district-selected measures of student or school performance, assuming they meet criteria established in state board rule and Arizona statutes and align with a teacher's instructed content area.

A summary of the types of assessments or data indicated by each state as appropriate for district consideration in supporting teacher evaluation is provided in column 6 of the table in Appendix 3. This table suggests that most states impose few, if any, requirements around the number and type of assessments that may be used to establish evidence of student achievement. In fact, in almost all cases, districts (and sometimes teachers) are able to select or propose any assessments or data that they believe will appropriately reflect student achievement in the skills or domain of interest. When states do impose requirements, they typically include: the use of data resulting from state or district administered end-of-course assessments (when available), the inclusion of one or measures of school performance, or the selection of assessments from a pre-approved state list. In some states, growth estimates for teachers in NTSG are based, all or in part, on school-based growth measures resulting from conditional status estimates of teachers in TSG (e.g., AR, FL, TX). In such cases, state standardized assessments are the primary source of data to support evaluation for teachers in both TSG and NTSG.

Although districts are typically given a great deal of control related to the selection of desired student achievement data for NTSG, review and approval of these SAM is required in some states. Approximately 1 in 5 states require some level of state approval of selected assessments or data prior to use, and of those states three-fourths were RTTT winners. In addition, states which propose the use of multiple SAM often dictate the use of a pre-approved assessment to establish one SAM, but leave the selection of additional sources of SAM to the district.

Given the flexibility afforded to districts, little information is available at this time about which assessments are most commonly used to support teacher evaluation in a particular NTSG. Gathering this information requires the collection of data at the local level, as it likely differs depending on a variety of factors, including available district resources and the types of assessments already in place. Subsequent research will explore the impact of these factors through case studies of select districts.

With respect to differences between TSG and NTSG, it is interesting to note that roughly two-thirds of states that dictate the use of state standardized assessments to inform teacher evaluation in TSG[11] do not require state approval of the assessments or tools selected for use in NTSG. While one may expect differences in this regard, the trends observed here suggest a much greater hesitancy on the part of some states to be as prescriptive in teacher evaluation for NTSG compared to TSG.

Analytic Approaches Used to Incorporate Measures of Student Achievement into Teacher Evaluation for NTSG

The previous section discussed the different tools and approaches states are using to obtain student achievement data in NTSG. This section summarizes the methods by which these measures are being used or proposed for use to support NTSG teacher evaluation. While a variety of analytic approaches have been considered to evaluate teacher effectiveness, the most common include growth models, conditional status models, goal setting models, such as those represented by an SLO process, and shared attribution. A brief summary of each of these approaches is provided in the table below.[12]

Table 4. Summary of Analytic Approaches

Analytic Approaches for Estimating Growth	Necessary Conditions	Procedures	Growth Interpretation
Growth Models	Pre-and post-test measures are available within the subject area of interest and exist on a common vertical scale	Calculate difference between pre and post-test performance on common scale	Gain (or Loss) in student performance between two points in time
Conditional Status Models (e.g., VAM-Models; SGP)	Pre-test data on one or more assessment(s) in or related to (i.e., correlated with) the subject are of	Condition on pre-test data (and potentially other covariates) as a means of evaluating	Difference between expected performance and observed

[11] See State of the States 2013 Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice at the link: http://www.nctq.org/dmsStage/State_of_the_States_2013_Using_Teacher_Evaluations_NCTQ_Report

[12] For a detailed summary of each approach and guidelines for use, see Marion & Buckley, 2011

	interest and subject-specific post-test data are available	post-test performance for a given student.	performance given prior performance in the same or a related subject area.
Goal Setting/ SLO Process	Process by which teachers use existing student performance data (of a variety of types) to establish learning goals for students in their class, and then evaluate student performance relative to those goals		Degree to which a student or group of students attained one or more specified learning goals
Shared Attribution	Any of a variety of techniques that involves the attribution of a common estimate of student growth, achievement or teacher impact on student learning — based on aggregation at the group, school or district level — to one or more teachers		Depends on nature of shared attribution approach

The first three rows in Table 4 outline approaches that allow for estimates of an individual teacher's impact on student learning. Growth and conditional status models typically accomplish this by calculating the mean or median student growth estimate for a given teacher and comparing it to a pre-established set of growth standards. Goal setting/SLO approaches, on the other hand, quantify a teacher's degree of attainment of the learning goals specified for his/her class relative to established growth expectations. Shared attribution does not provide for inferences about the impact of individual teachers, per se, but instead utilizes the performance of multiple teachers or a school to establish an aggregate indicator of impact deemed appropriate to support the evaluation of individual teachers.

In light of information provided by states, goal setting models and shared attribution (which can include shared SLOs in addition to shared VAM or SGP measures, gap reduction, graduation rates, etc.) appear to be the two approaches most commonly used (or proposed for use) by states. An SLO process, or something extremely similar to it, is discussed for roughly two-thirds of states, while nearly half of states suggest the use of some form of shared attribution either alone or in conjunction with another approach. Approximately 20 percent of states did not specify the use of a particular approach.

Even in states that share a general approach, the practical impact on students and teachers can differ substantially. For instance, the use of SLOs may differ considerably across districts in terms of who selects assessments, how many SLOs must be used, how weighting occurs, the type and level of performance required, and the amount of guidance provided to teachers and evaluators. The implementation of shared attribution is even more nuanced, as the level of attribution (grade, instructional team, school, district, etc.) as well as the metric used (e.g., math and ELA VAM scores, SGP, graduation rates, etc.) can completely change the meaning and consequences of shared attribution in teacher evaluation.

In practice, a variety of shared attribution procedures are currently used to inform the evaluation of teachers in NTSG, including:

- Attribution of conditional status results associated with TSG to NTSG: Teachers in NTSG are assigned the average teacher-impact rating for the district, school or a pre-defined group of teachers.
- Group-based goal setting: Common learning goals, or targets, are established for a specified set of teachers and the average rating earned is assigned to all teachers in this group.
- Attribution of school-based performance measures: A common measure of school performance is assigned to all teachers in a given school.

Of these shared attribution procedures, the first appears to be the most common, with approximately 1 out of 5 states proposing the use of conditional status results associated with TSG for NTSG. Of these states, a few are working diligently to establish assessment systems that provide for the calculation of student growth measures for all teachers. These include states such as North Carolina and Florida, which are developing end-of-course (EOC) assessments for every subject area, as well as Tennessee, which provides for the use of Teacher Work Portfolios as a means of establishing measures of student growth in select subject areas (e.g., Fine Arts and PE). For these states, the provision of additional measures and analytic approaches represents an ongoing process, such that the designation of NTSG actually changes from one year to the next.

While most states appear to promote the use of shared attribution and goal-setting approaches for NTSG, given the flexibility afforded to districts, it is unclear what is occurring at the local level. Approximately 60 percent of states indicated that the approach used to estimate student growth was either partially or fully determined by districts. Furthermore, 4 out of 5 states afford districts some flexibility to define the manner by which SAM are defined, aggregated, and weighted to establish a final growth or teacher impact rating in NTSG. Therefore, even if districts within a state share a general approach to operationalize SAM in teacher evaluations (e.g. SLO, shared attribution), the manner in which that approach is implemented may differ greatly from district to district.

When specified, the set of analytic approaches used to support evaluation in tested subjects and grades (TSG) and NTSG often differ within a given state. This is due in large part to the availability of standardized student achievement measures which allow for the use of conditional status models in TSG, as previously discussed. Based on our analysis, approximately 1 out of 3 states *explicitly* outlined the use of a value-added model or student growth percentile procedure in TSG. However, of the states which indicated the use of an SLO-type procedure for NTSG, less than half also prescribed this approach (either alone or in conjunction with conditional status approaches) for tested subjects. This is an important point given that SLO and other goal-type procedures are typically associated with higher degrees of local control than that afforded by conditional status approaches. For example, Maryland requires that student learning objectives inform 50 percent of a teacher's final rating for both TSG and NTSG, but Mississippi only requires the use of SLOs for NTSG.

Index of Local Control

The use of SAMs to support teacher evaluation in NTSG is a relatively new endeavor in policy and practice. As a result, there is a great deal of variability across states with respect to the degree of control afforded to LEAs in developing and implementing teacher evaluation systems for NTSG. To better evaluate the nature of this variability, we established the NTSG Local Control Index (NTSGLCI). Creation of the index involved identifying those questions (from the list presented in Appendix A) believed to elicit information about the level of control afforded

to districts in defining their evaluation systems for NTSG, and analyzing the extent to which they provided useful, non-redundant information about local control. Ultimately questions 6, 7, 12, 13, 15, and 17 from the protocol were quantified and analyzed for inclusion in the index.[13] Responses to questions 6, 7, 12, 13, and 15 were operationalized as 0, 1, or 2, with 0 indicating completely dictated by the state, 1 offering some amount of local control, and 2 giving full local control. Question 17 was operationalized dichotomously (0 or 1) because the practical implications of this item in terms of local control—whether or not a district has the ability to choose between a compensatory or conjunctive evaluation model—were deemed less important than the implications of the other five items. For each state, a NTSGLCI was created by summing the values assigned to each of these questions, resulting in score scale of 0 to 11. The states with a high NTSGLCI leave most/all decisions regarding the use of student achievement up to individual districts, including the structure of evaluation systems, how assessments are identified, and how scores are operationalized and aggregated. Conversely, states with a low NTSGLCI generally dictate how assessments may be selected, what analytic approaches are used, and how scores are incorporated into evaluation systems. Reliability analyses were conducted to test the internal consistency of this instrument, and a relatively high Cronbach's alpha[14] (α=0.8) suggests that the individual items perform reliably.

Table 5 shows the descriptive statistics for the six individual items, as well as for the NTSGLCI. The average NTSGLCI across all states was 7.6, with values ranging from 1 to 11. The most common NTSGLCI score was 11, which corresponds to essentially complete local control related to the specification and use of SAMs in teacher evaluation for NTSG. A total of eleven states received this maximum score, including states that have not received ESEA waivers, states that have yet to fully establish policy in this area, and states that strongly favor local control and have only issued suggestive guidelines around teacher evaluation. Despite the fair proportion of states with near-complete local control, considerable variation was exhibited.

[13]Question 9 was also examined for inclusion in this index. However, state responses to this question proved difficult to classify. Furthermore, reliability testing found that the operationalized response to question 9 decreased the internal consistency of the index variable. For these reasons, it was not included.

[14] Cronbach's Alpha is a measure representing how closely related a set of items are as a group

Table 5. Descriptive Statistics, Measures of Local Control in NTSG Teacher Evaluation (higher numbers indicate greater local control)

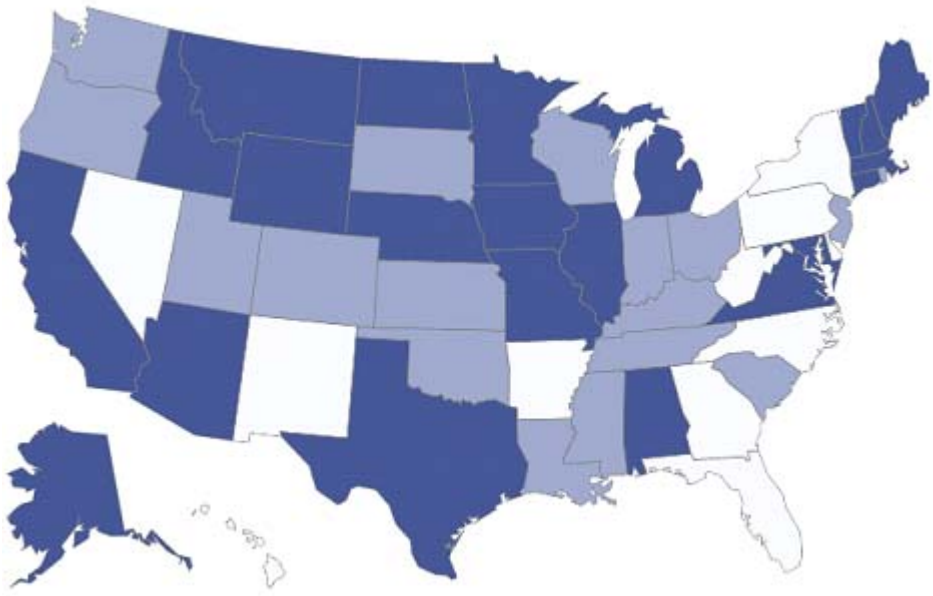
Abbreviated Protocol Question	Question Number	Range	Mean (RTTT States, Only)	Mean (non-RTTT States, Only)	Mean (All States)
Who Selects Measures of Student Achievement?	6	0 to 2	1.3	1.7	1.6
Who Approves Measures of Student Achievement?	7	0 to 2	1.3	1.7	1.6
Who Selects the Analytic Approach?	12	0 to 2	0.7	1.3	1.1
Who Aggregates Scores?	13	0 to 2	1.2	1.4	1.3
May Alternate Systems be Developed?	15	0 to 2	0.7	1.4	1.2
Who Determines Model Structure?	17	0 to 1	0.2	0.5	0.4
NTSGLCI		**1 to 11**	**5.5**	**8.1**	**7.6**

To clearly document the range of NTSGLCI, values were aggregated into three categories: high, moderate, and low local control: states with a NTSGLCI greater than 8 were deemed to have high local control; those with an NTSGLIC less than 5 were identified as low local control states; states in between were deemed moderate. Table 6 shows how category representation varies across RTTT and non-RTTT states. Figure 2 is a choropleth map of the US that shows which states fall into each category.

Table 6. Local Control in the Use of Student Achievement Measures in NTSG Teacher Evaluation, by RTTT status

	Number of States	Low Local Control (%)	Moderate Local Control (%)	High Local Control (%)
RTTT States	18	38.9	38.9	22.2
Non-RTTT States	32	12.5	31.3	56.3
All States	50	22.0	34.0	44.0

Figure 2: Choropleth Map Illustrating Low (Lightest), Medium, and High (Darkest) Local Control States in NTSG Teacher Evaluation in US



Both Tables 5 and 6 highlight different trends across states in light of RTTT status. Table 5 shows that states receiving RTTT funds exhibited lower degrees of local control, on average, across all individual measures than did the non-winning states. This trend is even more pronounced when looking across designations of overall local control in Table 6, with RTTT states being roughly three times as likely to preserve a low level of local control, and half as likely to allow a high level of local control.

Discussion

In response to incentives for state funding through RTTT and ESEA waiver requirements, states have become more deliberate about specifying plans for evaluating teachers in NTSG. While these plans are at different stages of completion, we identified several trends across states related to how NTGS evaluation systems are being defined and operationalized. Three of the most prevalent trends include: the use of SLO or goal-setting procedures to establish estimates of student growth; the use of some form of shared attribution to assign student growth or school/student achievement estimates to individual teachers; and the granting of high local

control to districts on the selection of assessment tools and/or data used as evidence of student achievement.

The latitude provided to districts specifically in selecting the source of student performance data used in NTSG begs questions as to: 1) the type and degree of support states are providing to districts related to identification of appropriate, high quality assessment instruments (or data) and 2) the extent to which states conduct or require local auditing of procedures used to collect and evaluate SAM for NTSG. While these questions were posed as part of our research, detailed information regarding these procedures typically was not readily available, suggesting this as an area where future investigation may be required. For example, while close to half of the states indicated some level of guidance is provided to districts to support the selection and/or review of measurement tools (e.g., through guidelines documents or online resources), details related to the nature of this guidance was typically not supplied. Furthermore, guidance materials, when specified, tended to vary greatly across states, ranging from assessment quality checklists to exemplars of high quality assessment tasks (e.g., banks of learning objectives and appropriate evaluation measures and lists of approved assessments and hyper-links to scholarly documents.

Similarly, while several states indicated that auditing procedures would be implemented to ensure efficacy in the teacher evaluation process, little if any detail was provided around how or when this would be accomplished — especially with respect to the collection and analysis of student achievement data. In the few cases where information was provided, state involvement in the auditing process included such things as the review of district plans for implementation, review and approval of SLO targets and objectives, monitoring of district adherence to state law and rules, and random monitoring of districts and schools around implementation. Some states indicated that auditing plans were still under development, or that only local auditing was required.

In addition to these overarching trends, we also identified several areas of variability across states including the:

- number and type of teachers that fall under the umbrella of NTSG;
- extent to which efforts are being put in place to move educators from NTSG to TSG status;
- emphasis given to school-based measures in teacher evaluation;
- degree of support provided to support districts related to the selection, development and review of high quality assessment materials and teacher evaluation procedures; and
- type and degree of local control afforded to districts in the design and implementation of their NTSG teacher evaluation systems.

In many cases, observed variability was related to RTTT status, as RTTT winners exhibited earlier deadlines to be fully operational, a greater emphasis on SAM, lower degrees of local control, and more detailed information regarding procedures used to support and evaluate the teacher evaluation process than non-RTTT winners. In other cases, variability was tied to a state's plans for establishing measures of student achievement and incorporating them into the NTSG evaluation system. For example, states pushing to establish EOC assessments in all grades and subjects were likely to have fewer educators associated with NTSG, with a goal of no distinction between TSG and NTSG in the future.

In most cases, we can only speculate as to the source of observed variability across states, as it likely relates to such things as the state's theory of action related to educator evaluation, the availability of certain assessments and data, fiscal and human resources, legislative constraints, and other factors that could not be established through a document review of the type conducted for this paper. For example, Pennsylvania includes the School Performance Profile, an overall index of school performance, as a SAM within the educator evaluation system for both TSG and NTSG because the state education department believes that its inclusion highlights the importance of "shared responsibility" at the school and student level and facilitates collaboration among educators — both of which are believed to support improved instruction and student learning. In future case studies with districts we will collect information

about the rationale and impetus for certain design decisions and the implications of those decisions on the evaluation process.

The high degree of local control afforded to districts in the implementation of NTSG is a compelling finding that highlights an additional factor that must be considered when discussing the similarity of teacher evaluation results across districts in a given state. While some states use SLO procedures and shared attribution approaches for both TSG and NTSG, the majority of states use conditional status models in lieu of or in conjunction with these procedures for TSG. Since such models typically necessitate the use of state-developed standardized assessments and complex statistical calculations to estimate value added, little local control can be afforded to districts around these aspects of the evaluation system. In fact, in most cases, value added estimates are supplied to districts based on state-specified calculations that operationalize the definition of "appropriate student growth" and result in a prescribed rating.

In contrast, the approaches utilized in NTSG provide for much greater local control. The SLOs process, for example, typically provides for local decisions with respect to the assessments or data used to provide evidence of student achievement, the manner in which identified SAM are used to estimate or evaluate student growth, the process by which multiple SLOs are aggregated, and the manner in which results are operationalized to provide for an overall rating of teacher impact to inform evaluation. Similarly, when shared attribution is employed, districts and teachers are often provided with a great deal of flexibility in determining the unit of aggregation (grade, school, and district) as well as the pieces of data to be considered (VAM results, other SAM).

We are not arguing for or against the provision of local control, as its impact will likely vary across systems depending on a variety of contextual factors such as the availability of local resources and the selected analytic approach. However, we do believe that significant differences in the control afforded to tested and non-tested subjects and grades for teacher evaluation could be a source of concern for stakeholders. For example, if teachers in TSG are evaluated using VAM while their NTSG colleagues use an SLO approach, teachers in tested subjects and grades may perceive they are being treated unfairly because they are being held

to rigorous state-defined standards, have less control over their evaluation, and are using controversial value-added procedures they don't fully understand. On the other hand, teachers in NTSG may have similar perceptions of being treated unfairly since they have to do additional work related to the specification and evaluation of SLOs, may not be adequately prepared to generate/identify the materials and data necessary to support the SLO process, and/or must rely on their evaluator to approve the quality of their performance related to student growth.

Whether such concerns are warranted is a question that can only be answered by additional research on the impact of different types and degrees of local control on desired outcomes, and the extent to which those findings interact with district, school and teacher-specific factors. It is our hope that the information in this document as well as the Non-Tested Subjects and Grades Local Control Index (NTSGLCI) can act as a stepping stone to support future work in this area.

References

Buckley, K., Marion, S. (2011). *A survey of approaches used to evaluate educators in non-tested grades and subjects.* Harvard University and National Center for the Improvement of Educational Assessment.

Doherty K.M., & Jacobs, S., (2013). *State of the States 2013: Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer New York.

Marion S. & Buckley, K. (2011). *Approaches and considerations for incorporating student performance results from "Non-Tested" Grades and Subjects into educator effectiveness determinations.* National Center for the Improvement of Educational Assessment.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. M., & Thorn, C. A. (2006*): The other 69 percent: Fairly rewarding the performance of teachers of non-tested subjects and grades*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Discussion paper.

Appendix A: State Questionnaire Related to NTSG

1. When is the system scheduled to be fully operational for teachers in NTSG?
2. Is there a state-developed model for teacher evaluation, or does the state supply teacher evaluation guidelines to districts?
3. What percent of a NTSG teacher's overall evaluation is based on student achievement measures?
4. What percentage of a Tested GS teacher's overall evaluation is based on student achievement measures?
5. What types of assessments, tools, or data may be used to establish evidence of student achievement in support of educator evaluation for NTSG (e.g., state summative assessments, locally developed classroom assessments, graduation rates, etc.)?
6. How are the assessments/tools/data used to establish evidence of student achievement in NTSG identified or selected for use? Who is responsible?
7. Does the state need to approve all assessments/tools/data used to support teacher evaluation for NTSG?
8. Do state developed procedures/guidelines exist to support the evaluation and use of assessments/tools selected to support educator evaluation in NTSG? (Provide examples if available.)
9. What combination of local and state oversight/auditing procedures are in place to support the collection and evaluation of student achievement measures for NTSG?
10. What analytic approach(es) are required or recommended to incorporate student achievement measures into teacher evaluations for NTSG?
11. Did this state receive RTTT funding? Did this state receive other funding which was specifically targeted at developing teacher evaluation systems in NTSG?
12. What entity is responsible for selecting the analytic approach(es) used to incorporate student achievement measures into teacher evaluations for NTSG?
13. What entity is responsible for determining how student achievement measures are defined, combined, and weighted to support the evaluation of teachers in NTSG?
14. Is the use of student achievement measures as part of the evaluation of teachers in NTSG dictated by law?
15. Does the state provide districts with option to develop alternate system with state review/approval? Explain if necessary.
16. Are the implications of evaluation the same for NTSG teachers and TSG teachers?
17. What type of model is used to assign educators an overall evaluation rating: a summative/compensatory model, conjunctive/decision matrix type model, or something else? Explain if needed.

Appendix B: Sample Completed State Questionnaire

NC Non-Tested Subjects and grades (Approved waiver draft date: May 2012

North Carolina Educator Evaluator System (NCEES)

Student Growth Overview: http://www.ncpublicschools.org/docs/effectiveness-model/resources/student-growth-overview.pdfhttp://www.ncpublicschools.org/docs/effectiveness-model/resources/measures-guide.pdf

1. When is the system scheduled to be fully operational[15] for teachers in NTSG[16]?	2014-2015? NC Final Exams in 2012-2013, K-3 in 2013-2014, Analysis of Student Work in 2014-2015. Effectiveness Statuses for teachers determined in Fall 2014, Fall 2016, and Fall 2017 respectively. Must be used to inform personnel decisions by 2016-2017.
2. Is there a state-developed model for teacher evaluation, or does the state supply teacher evaluation guidelines to districts?	Model
3. What percent of a NTSG teacher's overall evaluation is based on student achievement measures?	**1 out of 6 standards is focused on student growth**
4. What percentage of a TSG teacher's overall evaluation is based on student achievement measures*?*	**1 out of 6 standards is focused on student achievement** For this one standard: 100% Individual EVAAS Note: The state has expanded common exams to essentially to include all 4 core subjects in grades 4-12, or roughly 70 percent of all teachers in the state. Roughly 40 percent currently have a VAM, working on developing more. Student growth in subjects for which common assessments weren't created (e.g. Fine Arts) will be measured utilizing the Analysis of Student Work Process.
5. What types of assessments, tools, or data may be used to establish evidence of student achievement in support of educator evaluation	NC Final Exams were developed for teachers in traditional NTSG. These are essentially EOCs. Growth for K-2 educators is calculated

[15] For the purposes of this study, we consider "fully operational" to mean that a teacher evaluation system/model/guidelines is/are in place for all teachers.

[16] For the purposes of this study, we consider NTSG teachers to be those teachers who are not in math or ELA grades 4-8 and who do not have a state-supplied score for growth/value-added.

for NTSG (e.g., state summative assessments, locally developed classroom assessments, graduation rates, etc.)?	based on the Text and Reading Comprehension of mClass Reading 3D. Teachers in nontraditional NTSG (e.g. Fine Arts) are measured through the Analysis of Student Work Process.
6. How are the assessments/tools/data used to establish evidence of student achievement in NTSG identified or selected for use? Who is responsible?	All state-developed growth measures were determined through a process that involved teachers. These are then used for all teachers in the state in the subject area.
7. Does the state need to approve all assessments/tools/data used to support teacher evaluation for NTSG?	Yes
8. Do state developed procedures/guidelines exist to support the evaluation and use of assessments/tools selected to support educator evaluation in NTSG? (Provide examples if available.)	The state does provide guidelines and training to evaluators on how to implement the evaluation (observational) system for all teachers. NTSG assessments are in the development stages, but the state will provide guidance and procedures to districts when we begin full implementation (2014-15 SY).
9. What combination of local and state oversight/auditing procedures are in place to support the collection and evaluation of student achievement measures for NTSG?	The NCDPI monitors the evaluation of teachers and requires districts to justify cases where teachers did not receive a summative evaluation in the prior school year. Student achievement data is monitored through our roster verification process which requires teachers to "claim" instructional responsibility for the students who are registered for their courses.
10. What analytic approach(es)[17] are required or recommended to incorporate student achievement measures into teacher evaluations for NTSG?	Currently those teachers without VAM estimates have school-level VAM estimates as placeholder data in their student growth component of evaluation. School level data is not currently used to impact a teacher's evaluation. Analysis of Student Work employs a teacher portfolio approach to collect

[17] Analytic approaches refer to the different techniques used to incorporate student achievement measures into educator evaluation (e.g., VAM, SGP, Shared Attribution at a school or district level using VAM, SGP or other Measures; SLO approach, goal/target setting approach, conditional status modeling, teacher portfolio approach, etc).

	student growth information.
11. Did this state receive RTTT funding? Did this state receive other funding which was specifically targeted at developing teacher evaluation systems in NTSG?	RTTT Phase II Winner.
12. What entity[18] is responsible for selecting the analytic approach(es) used to incorporate student achievement measures into teacher evaluations for NTSG?	State.
13. What entity is responsible for determining how student achievement measures are defined, combined, and weighted to support the evaluation of teachers in NTSG?[19]	State.
14. Is the use of student achievement measures as part of the evaluation of teachers in NTSG dictated by law?	Per waiver requirements
15. Does the state provide districts with option to develop alternate system with state review/approval? Explain if necessary.	Yes—districts can opt out of certain requirements.
16. Are the implications of evaluation the same for NTSG teachers and TSG teachers?	Yes
17. What type of model is used to assign educators an overall evaluation rating: a summative/compensatory model, conjunctive/decision matrix type model, or something else? Explain if needed.	Conjunctive/decision matrix model. Teachers must show proficiency on five observational measures as well as the student-growth measure. A lack of proficiency in any one area will result in the teacher being rated as "In Need of Improvement".

[18] Entity here refers to state, district, school, or teacher/evaluator. If the entity is indicated as "local", it simply means "not the state".

[19] For example, who determines: the number of score or proficiency levels associated with an SLO, the level of aggregation associated with a shared attribution measure (class or school), the manner in which multiple student achievement measures are combined, etc.)

Appendix C:

Summary of State Responses

http://www.nciea.org/publication_PDFs/Table%20for%20Appendix%20C%20082814.pdf