A System for Using Student Academic Growth in the Evaluation of Teaching Effectiveness in the Non-Tested Subjects and Grades

A Guide for Education Policy Makers and Evaluators of Teachers
# Table of Contents

**Preface** .................................................................................................................. 3  
**Acknowledgments** ................................................................................................. 4  
**Student Academic Growth and Teacher Evaluations** ........................................ 5  
  - The Problem in the Non-Tested Subjects and Grades ............................................. 5  
  - What Are the Options for the Non-Tested Subjects and Grades? .......................... 5  
  - Test-Based Value-Added/Growth Indicators .......................................................... 6  
  - Student Learning Objectives (SLOs) ................................................................. 6  
  - Comparisons of Approaches ......................................................................... 6  
  - Human Judgment and Multiple Measures ............................................................. 9  
  - Interpreting Normative Data ................................................................................... 9  
  - Weighing Multiple Measures .................................................................................. 9  
**The Recommended “Simple Regression” Approach** .............................................. 10  
  - Common Assessments ........................................................................................... 10  
  - Why Not Simple Pre-Post Growth? ....................................................................... 10  
  - Simple Prediction/Regression ................................................................................ 11  
  - Numbers of Students and Teachers ...................................................................... 12  
  - Outcome or End-of-Course Measures .................................................................. 13  
  - More on Performance Components ....................................................................... 14  
  - Predictor Variables ............................................................................................... 14  
    - Predictors in General ......................................................................................... 14  
    - Predictors to Use ............................................................................................ 15  
  - Associated Analyses and Checks ......................................................................... 15  
  - Some Final Words on the Generation, Interpretation, and Use of Value-Added/Growth Statistics ................................................................. 16  
    - More on the Proposed Method ......................................................................... 16  
    - How Much Work Is It? .................................................................................... 17  
**Appendix A** .............................................................................................................. 18  
**Overview and Recommendations of the AASETE Study** ................................ 19  
  - Major Findings of the AASETE Project ................................................................. 19  
  - Recommendations from the AASETE Project ....................................................... 20  
**Appendix B** .............................................................................................................. 21  
**Instructions for Performing Regression-Based Growth Analysis at the Teacher Level Using Excel** ................................................................. 22
Preface

In 2011, the Bill and Melinda Gates Foundation funded the research and development project entitled Alternative Assessment Strategies for Evaluating Teaching Effectiveness (AASETE). The primary purpose of the project was to design a research-based system for using performance assessments along with other instruments to measure student academic growth, which in turn, could be used with other measures of teaching effectiveness for purposes of teacher evaluation in non-tested subjects and grades. These subjects and grades are those for which there are not state assessments that can serve as pre- and post-test measures of student achievement. This guide describes a recommended system and is intended for education policy makers and evaluators of teachers.

Sections of this guide discuss the political context surrounding teacher evaluations, the challenges and options in the non-tested subjects and grades, and finally the recommended system for districts to use in generating student progress indicators for teachers. This guide could have briefly presented the recommended approach; however, it was deemed important that policy makers understand (1) the issues associated with using student growth indicators for teacher evaluations and (2) the pros and cons of the various approaches, in order to make informed decisions.

The proposed system calls for districts or district consortia to collect common prior achievement data on students, administer common end-of-course assessments (including performance components), and apply a simple prediction model using standard Excel regression software. A teacher’s “score” or student progress indicator from this method is the average of his/her students’ differences between their end-of-course test scores and predicted end-of-course test scores.

Appendix A provides an overview of and recommendations from the research component of the AASETE project, the results of which, along with a literature review and the collective expertise and experience of a technical advisory committee, guided the design of the system proposed in this document. Appendix B is a worked example of the Excel analysis that could be used to implement the system.

This guide is not intended to be a formal report of academic research; rather it is intended to be a practical, readable resource for less technical audiences. Consequently, it does not cite the various sources drawn upon to inform the design of the AASETE study and to support the conclusions and recommendations presented herein. A separate document, Alternative Assessment Strategies for Evaluating Teaching Effectiveness: End-of-Action Report (Measured Progress, 2014), is a comprehensive report of the AASETE research effort. In it, the reader will find a list of references addressing important topics related to the use of student achievement results in the evaluation of teachers.
Acknowledgments

The AASET team at Measured Progress wishes to thank the members of the project’s steering committee for their wise counsel at several points during the conduct of the study. The measurement experts and state department personnel who served on this committee are identified below.

<table>
<thead>
<tr>
<th>Group Member</th>
<th>Affiliation</th>
<th>Group Member</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott Norton</td>
<td>Louisiana SEA</td>
<td>Joan Herman</td>
<td>CRESST</td>
</tr>
<tr>
<td>Fen Chou</td>
<td>Louisiana SEA</td>
<td>Lorrie Shepard</td>
<td>Univ. of Colorado</td>
</tr>
<tr>
<td>Rebecca Woosley</td>
<td>Kentucky SEA</td>
<td>Drew Gitomer</td>
<td>Rutgers Univ.</td>
</tr>
<tr>
<td>Angela Norlander</td>
<td>Colorado SEA</td>
<td>Henry Braun</td>
<td>Boston College</td>
</tr>
<tr>
<td>Glen Sirakavit</td>
<td>Colorado SEA</td>
<td>Brian Gong</td>
<td>NCIEA</td>
</tr>
<tr>
<td>Scott Marion</td>
<td>NCIEA</td>
<td>Elena Diaz-Bilello</td>
<td>NCIEA</td>
</tr>
</tbody>
</table>

Special thanks goes to Joan Herman and Julia Phelan of CRESST for their support in providing the performance tasks for the AASET history and life science tests. Thanks goes to Dr. Lisa Keller and the Abelian Group, LLC for their invaluable assistance with data analysis. Of course, the project involving large-scale testing, it is appropriate to acknowledge the Measured Progress staff in all departments contributing to the effort: Content Design and Development, Publishing, Operational Services, Scoring Services, and Research and Analysis. The AASET team consisted of Monique Ring, Jeanne Supple, Sara Bryant, and Stuart Kahl.

Of course, the study could not have happened without the tremendous cooperation of many individuals in the participating states (state department staff, local teachers and administrators, and students), who coordinated and conducted testing, provided data to Measured Progress, and actually took the tests.

The greatest appreciation is extended to the Bill and Melinda Gates Foundation whose commitment to education and support of teachers have been considerable. The foundation’s funding of AASET exemplifies its particular interest in teacher effectiveness. The support, patience, and wisdom of the Gates Project Officer, Ash Vasudeva, are especially appreciated.
Student Academic Growth and Teacher Evaluations

There is general agreement that teaching effectiveness is a, if not the, critical school-related factor in determining the quality and extent of student learning. Consequently, with concerns about the college and career readiness of our high school graduates and about the poor showing of U.S. students on international tests, education policy makers have moved the evaluation of teacher quality/effectiveness to the front burner. Two recent federal initiatives—the Race to the Top Program and the USDOE’s flexibility program (allowing NCLB waivers)—require that student growth, as evidenced by test scores, be weighed significantly in teacher evaluations.

Despite many challenges associated with the use of student test results in evaluations of teaching effectiveness, it is the job of teachers to raise the knowledge and skill levels of students; and so it would be foolish to assume that direct evidence of student learning can be left out of the process entirely. Some states and districts are already relying heavily on indicators based on aggregate student progress in this process, and even with some inherent problems with this practice, it might be assumed that it will see far greater use in the future.

The Problem in the Non-Tested Subjects and Grades

Tested subjects and grades are ones for which state tests can be used as pre- and post-tests for the computation of teachers’ value-added statistics. For example, with NCLB testing in mathematics at the end of grades 3 and 4, “growth” of fourth graders can be estimated using the Grade 3 state test as a pre-test or as a predictor of performance on the Grade 4 state test. But without a Grade 2 state test, Grade 3 mathematics would be considered a non-tested subject and grade even though there is a Grade 3 state post-test. Similarly, without a state mathematics test at Grade 9, that grade and subject would be considered non-tested even though there is a Grade 8 test for use as a pre-test or predictor. Of course, state test scores can still be used as predictors or as end-of-course tests in situations in which they are not available for both. Other subjects, such as science and history would be non-tested subjects at all grades in most states. Current literature estimates that approximately two-thirds to three-quarters of teachers are teachers of non-tested subjects and grades. Thus, the challenge in these subjects and grades is to identify and implement a viable alternative to the use solely of state test results for generating student academic progress indicators for teachers.

The focus of this document and the procedures it advocates pertain to academic subjects. To the extent that art and music, for example, are treated as academic subjects, the recommended system could apply to them. However, if in such areas, students are to produce pieces of art or perform musical numbers, it would be difficult to beat a portfolio of before and after work or performances by students as additional evidence of teaching effectiveness.

What Are the Options for the Non-Tested Subjects and Grades?

To be clear, the topic here is options relative to the use of student achievement data for teacher evaluations, with the full recognition that student achievement or growth indicators are only one of several measures that should be used to inform the evaluation of teachers. There are two general approaches to teacher effectiveness indicators based on student performance: “test-based” approaches and “student learning objectives” (SLOs). Treated as two distinct categories of approaches, these are not mutually exclusive as they can have some features, strengths, and weaknesses in common depending on how they are implemented.
Test-Based Value-Added/Growth Indicators

The value-added approaches that states are using in the “tested” subjects and grades and the approach recommended herein for the non-tested subjects and grades would be considered test-based. The term “value-added models” often refers to more sophisticated approaches to measuring student growth. However, this document does not make distinctions among such terms as “value-added scores,” “student growth/progress indicators,” “teacher effectiveness indicators,” etc. These all refer to the statistic computed for a teacher based on the implementation of any of a number of test-based approaches. Although different test-based models involve different statistical manifestations of student academic growth, they all involve the computation of some form of growth values at the individual student level and the aggregation of these at the teacher level, producing an effectiveness or value-added indicator for each teacher.

In the tested subjects and grades, if the state test results in different grades are put on the same vertical scale (a complex psychometric process), then the student growth value can be the simple difference between the current year state test score and the previous year state test score. If the state testing program does not use a vertical scale, then the previous year test scores can be used to predict students’ current year test scores, and a student’s “growth score” is the difference between his/her current state test score and his/her predicted score. (This is the approach recommended in this document for the non-tested subjects and grades.) In either case, a teacher’s effectiveness indicator is the average of his/her students’ growth scores. A third method, Student Growth Percentiles (oversimplified here), determines the percentile rank of each student’s growth score for a large population of students, and a teacher’s effectiveness indicator is the median percentile rank for his/her students.

Student Learning Objectives (SLOs)

Student learning objectives have been posited as an alternative to test-based approaches to evaluating student learning for purposes of teacher effectiveness evaluation, particularly in the non-tested grades and subjects. By this process, the individual teachers establish a limited number of learning goals for their students, monitor student progress toward these goals, and then evaluate the extent to which the students have achieved them at some point in time—perhaps at the end of a marking period or school year. Associated with an SLO is a metric or target for the teacher such as the percentage of his/her students achieving or exceeding a particular score on the measure of the SLO.

The creation of the SLOs, the development or identification of instruments to measure them, the decision rules about what level of student performance is required for students to be designated as having met each goal or objective, and the teachers’ targets/metrics relative to the SLOs are often left to the individual teachers, although such decisions would be monitored and approved by school administration.

The original intent of SLOs in a subject at a particular grade is the identification of, instruction in, and evaluation of performance with respect to four or five “big ideas” (major concepts or skills) within the discipline. SLOs look, smell, and feel like content standards adopted and implemented by states. In practice (several states have adopted the SLO approach), there is variation in the scope or “grain size” of SLOs, with some being very broad, calling for general proficiency in a subject area, and others breaking a domain down further. So too is there variation in the nature of the local instruments measuring them—more traditional commercial or teacher-made tests, performance tasks, and combinations thereof.

Comparisons of Approaches

Quality of student and teacher measures. State tests and commercial student achievement measures are generally considered to be of high technical quality. In terms of test reliability, that is almost certainly the case. Adhering to well-conceived content specifications, these instruments also represent targeted subject area domains well. (A qualification of that statement will be offered below in the section on domain coverage.) Thus, test-based approaches to teachers’ student progress indicators using such tests may have a greater likelihood of employing higher quality student achievement measures than approaches using district-developed tools and individual
teacher-developed measures. One might further argue that test-based approaches using district measures, selected or developed by teams of teachers, have a greater likelihood of using better instruments than SLOs if in the latter case, individual teachers are responsible for the development of the measures. Collaborative efforts often lead to better products.

It is very important to keep in mind that the reliability of teacher value-added or growth indicators is not at all the same as the reliability of the student achievement tests, the scores on which are the primary ingredients of those indicators. However, the real concerns expressed in the literature on growth models are matters of validity of the indicators, particularly in light of how they are used to make decisions about the relative effectiveness of teachers. A problem occurs when one tries to attribute the academic growth (or lack thereof) of a particular group of students solely to an individual teacher in spite of the myriad relevant factors, over which the teacher has no (or at best partial) control. Such factors include parental valuing of education, student motivation, peer effects, instructional support, school leadership, and many other contextual factors that vary across groups of students, making comparisons of teachers’ student growth statistics challenging. Even within a single school, where many of these factors are “controlled,” student differences exist across classes and teachers of the same subject and grade in part because assignment of students to classes is generally not random or otherwise “equitable” in terms of student characteristics, experience, and abilities.

An additional concern is that, as explained previously, there are many different ways of viewing and computing student growth, and different growth models produce different results in terms of the rank ordering of teachers based on their value-added scores or indicators. Further, it is well established and empirically verified that value-added scores for teachers are inconsistent not only across models, but also across years. While aggregating teacher data across years can reduce variability due to differences among the groups of students any particular teacher is assigned, this does not control for the many other factors affecting teacher value-added data that make comparisons of such results across teachers especially challenging. This does not mean, however, that a teacher evaluator should not look for evidence in value-added data that a teacher’s students have learned something. The real challenge relates to how the data are interpreted and used.

Thus, with respect to test-based value-added/growth models, there is general concern about whether they yield measures of the relative effectiveness of individual teachers or something else. The SLO approach is intended to minimize this concern by focusing on each individual teacher’s agreed-upon “deliverables” —student performance relative to the teacher-specific SLOs and target student performance metrics. Certainly during the school year, the individual teacher has a lot of control over these elements. However, there are clearly some offsetting disadvantages to the SLO approach as well— disadvantages due to the extreme “localization” of the information the SLO approach generates. Various aspects of validity, both with respect to test-based growth models and SLOs are further discussed in the next three sections.

**Domain coverage of student achievement measures.** During the course of a school year, a teacher “delivers” a full year’s curriculum. Thus, it makes sense that an end-of-year measure of student performance (or an aggregate of major interim summative measures) for use in teacher (and program) effectiveness evaluations should provide good coverage of that domain. This is content validity.

That state assessment instruments adhere closely to content specifications built to cover the state content standards is another reason state tests are considered high quality. However, teachers often express concerns that such tests don’t cover a lot of what they teach. While their concern may be based on a lack of understanding of sampling from a domain, if they believe major sub-domains are not covered, their criticism may be more justified. For example, current NCLB testing in most states focuses on efficiently measured, isolated knowledge, and low level skills, and not deeper learning. Even the upcoming next-generation tests of the major state assessment consortia, while they will do a better job of addressing deeper learning, will still be limited because of time and security considerations associated with high-stakes, on-demand tests.
It seems appropriate that tests to be used in evaluating programs and teachers would give reasonable attention to the broader goals of education, including deeper learning. Thus, in the non-tested subjects and grades, local districts have the opportunity to remedy a shortcoming in state tests by assuring that their own measures give adequate attention to higher order skills by including performance components that are extended tasks requiring students to demonstrate their abilities to apply foundational knowledge and skills. Despite old myths about performance tasks, they can address important curricular topics and skills, and they can do so reliably.

Because district tests are generally products of teams of educators, there may be a higher likelihood that those tests do a better job of covering a subject area domain than an individual teacher’s instrument addressing the SLOs uniquely identified by the teacher. Again, as is the case with test reliability, this is a matter of likelihood and not what has to be. Further, in practice SLOs and associated measures are not always totally unique to individual teachers. Such situations exemplify a gray area between the test-based and SLO approaches.

**Instructional impacts.** A major argument for SLOs is that the involvement of teachers in the processes of creating the SLOs, measuring student performance relative to them, and setting performance targets assures teacher ownership or “buy-in” and a commitment to address them effectively and monitor student progress toward them. Thus, the claim is that the SLO approach is geared toward instructional improvement and not punitive accountability and accountability measures “from on high.” However, with respect to test-based approaches to student growth in the non-tested subjects and grades, teachers (and schools and districts) can certainly collaborate to develop common student assessments, thereby creating some degree of teacher ownership.

SLO advocates further remind us of the common concern about high-stakes state testing leading to a narrowing of the curriculum in order to raise test scores. However, SLOs could also have such a negative impact. As suggested in the previous section on domain coverage, the nature of the SLOs is critical. The literature speaks of teachers, early in the school year, identifying particular weak areas for their students and creating SLOs that address them. However, if the SLOs do not reflect good coverage of the year’s curriculum, then teachers could give greater attention to SLO topics and skills and neglect others, thereby assuring good evaluation results, while not being particularly effective with respect to those other topics and skills. Of course, school administrators should be knowledgeable of these issues and guard against this possibility when approving SLOs. Nevertheless, the heavy involvement of teachers in deciding on SLOs and associated metrics is no guarantee that there will be no concerns about the “foxes guarding the henhouse” and no “gaming of the system” or negative instructional impacts.

**Normative information versus pre-determined target standards.** Student test scores are only meaningful in comparison to other data points—scores of other students (normative data), an established score standard, or previous scores. This statement can also be made for teacher effectiveness indicators based on student achievement information. In using student progress data for evaluating teaching effectiveness, the test-based approaches tend to rely on normative data (indicators for many teachers), while SLOs tend to rely on pre-established effectiveness indicator standards or targets. Of course, in either case over time, previous years’ indicators can be helpful in monitoring improvements in teaching effectiveness.

Test-based approaches commonly lead to a rank-ordering of teachers or judgments about teachers based on where their effectiveness indicators fall relative to those of others. This is both a strength and a weakness of the test-based approaches. It is a weakness because of the validity/comparability issues discussed earlier, but it is a strength because with knowledge of a norming group or subgroup of teachers, the normative data can help an evaluator better understand how much growth is reasonable or can be expected.

In establishing and approving metrics associated with SLOs (e.g., a target for the percentage of students scoring at or above a particular score on a test), the teachers and administrators, at least initially, may be flying blind—with no idea of whether a target is too low, a reasonable stretch, or too high. Adjusting the targets based on the results of the particular teacher could run the risk of facilitating continued low effectiveness. Also, adjustments to targets made in subsequent years would require that the SLOs and measures of performance relative to them remain unchanged.
Human Judgment and Multiple Measures

Interpreting Normative Data
Since the approach to using student growth for teacher evaluations recommended in this document is a test-based approach, further discussion of the use of normative data is appropriate. The evaluation of an employee is the responsibility of his/her immediate supervisor. In the case of teachers, with the issues associated with value-added statistics and other indicators of effectiveness, it makes sense that a teacher’s evaluation should be the result of the supervisor’s human judgment—judgment informed by data, not driven by it. Only the principal or department head can evaluate the data, taking into account the unique situation of the teacher—the student characteristics, school contexts, and other factors. This is the case even in situations in which a common assessment is taken by the students of many teachers of the same grade and subject, teaching to the same content standards. That a teacher’s value-added score is lower than those of others needs explanation, and that explanation, often in terms of the group of students served or other contextual factors, leads to a fairer judgment about the teacher’s performance as well as to possible interventions. Even though there are comparability issues with teachers’ value-added scores, normative value-added data (results for identified groups of teachers and the students they serve) can still be useful to a teacher evaluator to establish what might be reasonable levels of student growth, providing a basis of comparison, differences from which can be investigated and perhaps explained, even justified.

Regarding what teacher value-added scores should be compared to those of others, some value-added advocates argue that averaging three years of results would produce a more stable measure to use. While this is true, the practice assumes there is a need to reduce the teacher’s value-added statistics to a single number. This practice may be most appropriate for research purposes in which teachers’ scores are considered in aggregate. However, if human judgment is relied on more than a formulaic approach in teacher evaluations, then the immediate supervisor can view the three years of data separately and explain variations that may be due to factors other than statistical “noise.”

Weighing Multiple Measures
Just as accepted standards of testing practice state that important decisions about students (e.g., promotion or graduation decisions) should not be based on results of a single measure, so too does this principle apply to decisions about teachers. In the case of teachers’ value-added scores or any single test-based indicator, abiding by this rule is even more important for the simple reason that even though the tests scores of students may be highly reliable and can lead to valid decisions, the teacher effectiveness statistics that use them have issues, as described previously, that make drawing conclusions about a teacher’s effectiveness difficult.

Fortunately, there is general recognition that multiple and varied measures are needed for the evaluation of teachers. In addition to student achievement gains, such measures as classroom observations; parent, teacher and student surveys; and teaching artifacts (e.g., lesson plans, tests, student work) can be used in teacher evaluation systems. While the use of multiple measures is clearly a “must,” making sense of the resulting body of evidence is not without its challenges.

One of the results of the federal programs requiring that student achievement growth be counted significantly in teacher evaluation systems was that states passed laws assigning specific weights (percents of total evaluation) to student growth. Weights of 40 or 50 percent were not unusual. There are two problems with this. First, those are very high percentages for measures with limited comparability across teachers. Second, the assigning of specific weights is often associated with formulaic approaches to obtaining a final score for a teacher requiring the quantification of all the measures used, all of which may have issues with respect to reliability and validity. This practice encourages data-driven decisions about teachers, rather than data-informed judgments that can take into account the varying conditions and contexts under which teachers operate. Additionally, it is not clear that the weights assigned to the multiple measures are the actual weights they receive. If, for example, there is greater variation in the student growth scores of teachers in comparison to the variation in the other measures, the actual weight assigned to student growth could be considerably greater than the intended weight.
It seems that in this country we have come to rely far more heavily on formulaic, quantitative approaches than on human judgment in evaluations of people and programs, assuming greater fairness is a benefit of these approaches. This is not necessarily the case. It is ironic that we want to remove human judgment from important decisions adults must make at the same time we want to do a better job of developing higher order thinking skills in our students. If there is an issue with supervisors’ abilities to evaluate employees, then that issue should be dealt with directly.

The Recommended “Simple Regression” Approach

The AASETE recommendations call for teachers, schools, or districts to join together to use common student assessments (including performance components) and apply a simple prediction or regression model using raw post-test or end-of-course test scores and common spreadsheet software such as Excel. A teacher’s student growth indicator from this method is the average of his/her students’ differences between their actual and predicted end-of-course test scores. The method can also be applied in a Student Learning Objectives (SLO) environment if shared SLOs and common assessments are involved, as well as in the tested subjects and grades. Nevertheless, this is a “test-based” approach chosen because of (1) the greater likelihood of higher quality measurement resulting from collaboration of teachers, schools, or districts in test development, (2) the normative information it generates to assist in conclusions about teaching effectiveness, (3) its pre-empting of concerns about bias due to teachers having too much control over the metrics against which they are to be evaluated and undesirable practices to which that can lead, and (4) the SLO approach’s lack of a basis for identifying SLO targets. Additional rationales are provided in the more detailed discussion of the approach below. Appendix B presents a worked example of the approach, showing Excel screen shots with results of the required analyses and providing discussion of those results.

Common Assessments

As explained earlier, despite the challenges associated with comparing teacher growth indices across teachers, classes, grades, years, etc., there is utility in the use of normative information on teachers’ student growth indicators. It provides a basis for human judgments by the teacher evaluator about teachers’ performance in terms of student achievement. It establishes a norm, deviations from which need to be explained and justified (or not) by the evaluator taking into account the many factors that affect comparability.

Clearly, the achievement measures administered to students are central to the use of student growth in the evaluation of teaching effectiveness. If a test-based approach is used, common assessments (or equivalent tests) are a must, even though there would still remain factors affecting comparability of teacher effectiveness indicators. Thus, comparisons should be restricted to teachers of the same subject at the same grade level and addressing the same curricular standards. Common assessments are what enable the generation of normative test results. For the recommended system, both end-of-course tests and predictors must be common—the same for all teachers whose students are used in the proposed analyses.

Why Not Simple Pre-Post Growth?

One problem with the pre-post growth approach is that in many instances, appropriate pre-tests might not be available. Also, the nature of pre-tests is often an issue. Is it reasonable to test students on content to which they have not been exposed? If a test of pre-requisite knowledge is used as a pre-test, can the pre- and post-test results be placed on the same (vertical) scale? Also, a pre-test/post-test approach used for high-stakes teacher evaluations can be more easily corrupted or susceptible to “gaming the system” by teaching to the post-test. (Typically, a pre-test and post-test would be the same test or different forms of the same test.) Of course, telling students a pre-test is just to find out what they know at the start of a year and will not count could lead to inappropriately low scores. The use of a predictive or regression model opens up many possibilities in terms of predictor measures that can be used.
Predictors would not be nearly as limited in terms of content as a pre-test would be. What is required is correlates of end-of-course tests, not equivalent or equated tests.

**Simple Prediction/Regression**

In the tested subjects and grades, complex psychometric techniques are typically applied for scaling and equating tests, and sophisticated value-added models for analyzing student growth are used for the purpose of teacher evaluation. The AASETE study suggests that little is lost by the use of less sophisticated approaches. Thus, the recommended approach herein uses routine multiple linear regression analysis that can be run from common spreadsheet software such as Excel.

The regression approach is sometimes referred to as a prediction approach. In some applications of predictions, based on data both on predictors and an outcome or dependent variable, a function of the predictors that best estimates outcome scores is determined and then used in the future when only predictor information is available. In the teacher evaluation context, however, we are not looking into the future, but rather determining that relationship or function, and then computing predicted student end-of-course scores to which we compare the actual end-of-course scores that were already available and used collectively to determine the relationship. (This process is just the simplest case of the more sophisticated Value-Added Model or VAM.) The graphic below illustrates the process.
Assume there are hundreds, perhaps even thousands of green dots in the scatterplot, each representing a student. The students are from classes taught by many different teachers. The regression analysis determines the best combination of predictors, which produces the “narrowest” oval pattern. Also, it then determines the line of best fit through the oval. As an example, Student A’s “score” on the predictor combination is that under the arrow on the horizontal scale. That student’s actual score on the end-of-year test is shown by the tip of the top arrow on the vertical axis. However, the best prediction for students scoring the same as Student A on the predictor(s) would be the one at the tip of the bottom arrow on the vertical axis—determined by the line of best fit. In this case, Student A scored higher than was predicted.

For every student linked to a particular teacher, the difference between his/her actual and predicted end-of-course scores (the distance between the two red dots) can be computed. Averaging these differences (sometimes called residuals) across all that teacher’s students yields a measure that can be considered a student growth or progress indicator for that teacher. Doing the same for all the teachers whose students are represented produces a distribution of teachers’ indicators which shows that some teachers’ students on average do better than predicted and other teachers’ students do not. (A typical teacher’s indicator would be near zero.) As explained in earlier sections, this does not mean that the latter teachers are not effective. Their students may well have made progress, but not as much as was predicted. It is up to the teacher evaluator to judge if there is something about those teachers’ students or other factors that justify the lower teacher scores.

That’s enough on statistics, regression, prediction, etc. Once the scores of students on predictors and end-of-course measures are provided to Excel, the software does the work. In the next few sections, sample sizes, end-of-course measures, and predictor variables are discussed.

**Numbers of Students and Teachers**

To implement the “simple regression” approach, one needs to address the issue of the numbers of students and teachers required. The number of students may be less problematic. Considering a situation in which there are only two teachers of the same subject and grade involved, if each teacher is responsible for 120 students, then there are 240 students whose academic performance data can be used in the regression analysis. Very likely, there is enough variability in the students’ predictor and end-of-course scores that the necessary relationship between them can be established and used to implement the system. Thus, it remains to consider the number of teachers needed for meaningful interpretation of normative results.

In the tested subjects and grades, of course, there would be hundreds, even thousands, of teachers in a subject at a grade for whom value-added scores can be computed based on state tests. If a state mined the data, it could produce aggregated results for subgroups of teachers representing different populations served relative to types of communities, percent of students receiving free and reduced lunch, etc. Such data could be useful in helping teacher evaluators understand the impacts of some of these factors on the scores.

The situation in the non-tested subjects and grades could be more challenging. If collaboration among districts is a consideration, like districts in terms of populations served would be desirable. On the other hand, let’s consider our extreme example of a school with two teachers of the same subject at a grade. Suppose one is an experienced master teacher whose capabilities the principal knows well based on years of seeing the teacher’s assignments, tests, and student work products and other information on student performance. And suppose the second is an inexperienced teacher, but whose students are generally as capable as those of the master teacher based on prior academic performance. If the master teacher’s student results place most of them above the diagonal in the graphic on page 17, while those of the inexperienced teacher are mostly below the diagonal, that is useful information. It does not mean necessarily that the inexperienced teacher is doing a bad job, but it does suggest that greater student performance can be achieved in the future.
There are no magic numbers indicating how many teachers are enough. Statisticians will say simply that more are better. Would a dozen or fifteen be desirable? Sure, but the example described above suggests that the real need is for the evaluator to have a good understanding of the “norming group” available—that is, understanding the population(s) of students served and other contextual factors the teachers face.

Outcome or End-of-Course Measures

State tests used as pre- and post-tests or predictors and end-of-course tests for generating growth measures are generally of high quality technically. In addition to being highly reliable, considerable attention is given during their development to alignment to and coverage of state content standards. At the same time, because of time and cost constraints, these measures are sometimes criticized by educators for not covering important knowledge and skills they address in their instructional programs. In designing systems for evaluating student growth in the non-tested subjects and grades, educators actually have an opportunity to correct this common shortcoming of the cost-efficient, high-stakes tests. However, locally developed tests are often criticized for a lack of technical quality. It is important that any tests contributing to important decisions about students or teachers be of high quality.

The tests developed for the AASETE study were designed to be adequate for research purposes. That is, their reliabilities approached those of state tests. Although minimal in their coverage of content standards, they were carefully aligned to state standards; they included a significant performance component; and their results demonstrated pre-post growth during the school year. Nevertheless, in implementing student testing for actual teacher evaluation purposes, one would want more comprehensive coverage of course content and performance components more aligned to discipline-specific standards (or perhaps Student Learning Objectives).

Generally, a minimum of approximately 50 score points are required to produce an acceptably reliable test. (More would be better for a more comprehensive instrument providing better coverage of course content.) Those points can (and should) be achieved by a combination of efficient, one-point selected-response items as well as multi-point constructed-response questions and performance components. A more comprehensive measure can be obtained if instead of scores on a single end-of-course test, the sum or average of each student's scores on common end-of-term or marking period tests are used. Student work from common curriculum-embedded performance assessments can also contribute.

Fortunately, there are resources that can be especially helpful in assuring that the “locally developed” measures used for end-of-course tests are of high quality. First, item banks from test publishers (and perhaps from state assessment consortia in the future) can be used to supply many of the needed items and tasks. Second, since common assessments at some level are recommended, teachers, schools, and districts working together on end-of-course tests means there can be pooled resources, including existing test items and tasks, and monies for additional items and tasks. Additionally, partner schools or districts could more likely afford shared services of consultants or even regular positions for the development of tests and analysis of results. Publishers’ off-the-shelf tests may be of use here, but the same concerns educators have about state tests might make these instruments of greater use as predictor measures.

If student tests are to be used to contribute to teacher evaluations, then security issues and scoring bias have to be addressed. Regarding the first, although teachers from across districts can be involved in assembling item and task banks, it would be advisable for a testing specialist (not a teacher in the grade and subject for which tests are being developed) to be responsible for final test construction according to previously established specifications and that new test construction be done every few years, if not every year. As for scoring bias, there are online, distributed scoring systems that can allow the teachers to be involved in the scoring of student responses to constructed-responses and performance products, but at the same time prevent them from scoring the work of their own students. Double scoring (scoring of each student response by two readers independently) can help assure the quality of the scores. For non-selected-response components, there is nothing that says teachers cannot score their own students’ work for their own immediate uses and still submit the student work for distributed scoring.
More on Performance Components

The special need to generate data for use in evaluating teachers in the non-tested subjects and grades represents an opportunity—the opportunity to produce better measures than those used in the tested subjects and grades, through the involvement of teachers in a collaborative process that also gives the teachers a sense of ownership in the process and the instruments. The improvements in the measures could include greater attention to deeper learning through the performance assessment components.

Performance tasks ask students to apply foundational knowledge and skills in solving real problems or otherwise creating some form of product, presentation, or demonstration. In ELA, science, and social studies, students could be asked to write essays in the explanation or argumentation modes drawing supporting ideas for positions from literary or informational readings. In mathematics, a performance component could be a set of challenging constructed-response problems, each focusing on a different subdomain of mathematics—geometry, statistics, etc. These are just a few examples of tasks for on-demand testing. Sample tasks provided by the Smarter Balanced Assessment Consortium and PARCC provide other examples.

Curriculum-embedded performance assessment refers to instructional units that include multiple learning and evidence-gathering activities, some of which may lead to products or performances that are evaluated for formative purposes, and some that are scored for summative purposes. The latter could well contribute scores to the “outcome measure” used in the simple regression approach to student growth recommended herein.

Performance assessment’s time has come. Two decades ago, common concerns about performance tasks pertained to their inattention to important content and to low technical quality often related to their scoring. These need not be concerns today as high quality tasks and scoring approaches are available for use, and to serve as models for local development and implementation. It is through performance assessment, both on-demand and curriculum-embedded, that we can facilitate deeper learning and its measurement.

Predictor Variables

Predictors in General

Predictors need to be positively correlated with the outcome variable or end-of-course test, but they do not need to be as closely aligned to the standards or objectives covered by the outcome measure as a pre-test or equivalent form. In fact, they need not even address the same content domain. A little more low-level discussion of statistics follows.

Another way to think about the strength of predictors is in terms of the amount of variance in the outcome measure they account for or “explain.” Think about the scatterplot discussed two sections ago. There is a correlation coefficient representing the correlation between the linear combination (a function) of the predictors and the outcome measure (end-of-course test). The square of that correlation coefficient is the proportion of outcome variance accounted for by the predictors. The higher the correlation is, the greater the accounted-for variance is, which means the narrower the oval pattern in the scatterplot is or the more closely clustered around the line of best fit the students’ actual end-of-course scores are. Of course, we’re counting on the teachers accounting for much of the remaining variance.

As is commonly the case in regression analyses, the AASET study found that once two or three predictors were used, additional ones did not account for significantly more outcome variance. Thus, this tells us that in designing a system for generating our teacher value-added statistics, we should be using two or three predictors. This does not mean that other ones are not good predictors—it is just that much of the outcome variance they account for is already accounted for by the first two or three we use. This is because of their correlation with the predictors used.

For a variety of reasons, it is recommended that only academic variables be used for predictors. We know variables like socioeconomic status and race are correlated with academic achievement, for example. But that means they
have impacted the scores on academic predictors already, so some of their impact is already taken into account in our regression approach anyway.

**Predictors to Use**

The AASETE predictors of end-of-course test scores, in order of predictive value, were same-subject pre-test scores, same-subject state test scores from the end of the previous year, other subject state tests, and same-subject previous course grades. Since we recommend against pre-post change as a growth measure, we view the AASETE pre-test as another high quality predictor test much like a state test from the previous spring. Thus, a must-have predictor should be a same-subject test of demonstrated high technical quality. Match to specific course content is not as important for a predictor as it is for an outcome or end-of-course measure. State tests administered at the end of the previous school year or off-the-shelf commercial tests administered then or at the start of the year would fit the bill.

Next on the list in terms of predictive value would be a **high quality test in another subject**. For social studies or history, an English language arts test (reading and/or writing) would be a good second predictor. For science, a mathematics test would be a good bet. Again, previous year state tests or commercial products are good candidates.

A recommended third predictor would be **previous course grade in the relevant subject**. Despite the lower correlations between previous grades and outcome measures due to inconsistent standards of performance and grading practices across classes or schools, grades tap something that allows them to improve predictions of end-of-course performance somewhat. (Note: For the regression analyses, letter grades should be converted to numerical scores, and the same or similar numerical scale should be used across sites. The Wesleyan University website [http://www.wesleyan.edu/registrar/general_information/grade_chart.html](http://www.wesleyan.edu/registrar/general_information/grade_chart.html) provides a table for converting letter grades to a 4.0 scale.) While another predictor or two would do no harm, they likely would not improve predictions significantly if three predictors such as those described above are used.

**Associated Analyses and Checks**

Since the recommended approach involves student testing, it is important to know that the tests (predictors and outcome measures) are of high quality. Their reliability coefficients should be 0.75 at a minimum. (A 50-point test will generally exceed that minimum reliability as long as the item quality is reasonable.) As suggested above, the content of predictors is less critical than their predictive value, but the outcome measures should exhibit strong content validity. That means they should provide a good sampling/representation of the content domain of interest—the content standards and related curriculum. Test specifications, assuming they are met, are most helpful here—the spread of items/tasks across content categories and cognitive process levels (depth of knowledge) is important.

As with any data analysis effort, checks of the quality of the raw (input) data are a must. In addition to spot checking individual students' records to be sure the correct data (variables) are being used, routine descriptive analyses should be run to be sure only valid values appear in the data file for all students for all variables.

A worked example of the regression analysis needed to generate the teacher effectiveness indicators based on student achievement data is provided in Appendix B. The multiple regression analyses run on the student data should provide evidence of effective prediction of outcome or end-of-course scores. An R-squared of 0.5 to 0.6 should be a target. That is, the predictors should “explain” 50 to 60 percent of the variance in the outcome measure. Another critical guideline for this analysis pertains to the number of teachers whose students are to be included in the regression analysis. This is the number of teachers making up the group of teachers whose student growth statistics will be compared. These are the teachers who would have to be administering a common end-of-course assessment—7th grade social studies teachers in a couple of collaborating schools or districts, for example. (The same predictor data must be available for all their students, as well.) While there is no magic number of teachers, 12
to 15 might be a reasonable minimum. However, as explained in an earlier section, a good knowledge of the teachers in a smaller group could enable effective interpretation of comparative results for those teachers.

With respect to the teacher “scores,” the average of the students’ deviations from predicted outcome scores, there are two concerns. First, while easily computed for any teacher, the statistic for teachers of small numbers of students should not be taken as a good indicator of student growth for those teachers. Clearly, those figures could be quite variable from year to year. While there is no specific minimum number of students for whom a teacher’s student growth statistic can be computed, a district’s teacher evaluation program might establish a number of students, such as 20, for which the teacher statistic is either not computed or asterisked if that number or greater is not attained. For many teachers in the middle or high school grades, the numbers of students should not be a problem. For an elementary teacher teaching a lesser number of students all subjects, a growth statistic could be computed separately for each subject and an average of them computed. (Different results for a teacher in different subjects would be useful to know about, too.) Of course, multiple years’ growth statistics should be monitored for all teachers.

Another concern, and one that should be checked after any analysis of data has been done, is the reasonableness of the results. Based on his/her knowledge of the teachers, teaching assignments, and students, a teacher evaluator should verify that the results make sense. This second issue also gets at the heart of the problems associated with value-added statistics—their interpretation. Judging the reasonableness of the results for particular teachers is a difficult judgment indeed. Building a sense of what is reasonable over years of observing the data is important, but challenging. States can be helpful in this endeavor by reporting how statistics of teachers serving different populations of students in the tested grades and subjects generally compare, but this requires significant data coordination and analysis at the state level.

The effectiveness of a teacher should not be considered known when a value-added statistic is generated. As with any test results, these statistics need to be investigated, understood, interpreted, and explained. The teacher evaluator needs to apply human judgment teacher by teacher to determine each teacher’s effectiveness based on multiple indicators of effectiveness and the evaluator’s knowledge and experience. With respect to the student growth indicator, he/she should take into account factors over which the teacher has no control, and to do so should understand the make-up of all the teachers involved in the statistical analysis and how others fare who serve similar populations of students as the teacher in question. Also, as mentioned previously, multiple years’ worth of data should be examined before decisions of any consequence are made.

Some Final Words on the Generation, Interpretation, and Use of Value-Added/Growth Statistics

More on the Proposed Method

The approach to evaluating student growth in this document, out of necessity for the non-tested subjects and grades, makes use of “local” common measures and an analytic method that can be accomplished with readily available software usable by individuals who may not be statistical or psychometric experts. This is not to say that the results of the procedures are any less believable than those of more sophisticated methods typically applied in the tested subjects and grades. Sophisticated Item Response Theory (IRT) scaling used in statewide assessment programs can achieve somewhat greater precision in individual student scores and is useful for equating different test forms or equating tests across years. But because of our focus on teacher results in a particular year, student test results are used in aggregate and random error therefore “balances out.” Additionally, it should be pointed out that a growth model used frequently in the tested grades and subjects, the Hierarchical Value-Added Model, is just a slightly more sophisticated regression approach than the one proposed herein.

Given that a teacher’s job is to help students achieve at higher levels, it is reasonable to gather direct evidence of every teacher’s level of success in that endeavor. However, one piece of evidence is often not conclusive, and considering the limitations of teacher value-added statistics, this is definitely the case in the area of teaching
effectiveness. Not only must this evidence be evaluated along with evidence from several other teacher measures, such as that gained from observations, student surveys, etc., it must be scrutinized very carefully itself. That is, one must not consider a teacher’s score the end product of a student growth analysis. A human judgment about this evidence is needed, but not before many of the factors affecting the teacher statistics are considered. How does this teacher’s students compare with those of other teachers in the analyses? How does this teacher’s result compare with those of others serving similar groups of students? Is this year’s result for the teacher an anomaly, or is it similar to the results for previous years? The evaluator should keep in mind, too, that the procedure yields a rank ordering of teacher results—yet, it may be that no teachers in the group are doing a poor job.

The training of evaluators and auditing of these human judgments necessary to assure the judgments are good ones are topics that go beyond the scope of this document. Yet given that student academic progress data constitutes the only direct evidence of how much students are learning, it is important that significant attention be given to the development of teacher evaluator skills, including the interpretation and use of teachers’ student growth indicators.

How Much Work Is It?

To generate student growth statistics for all teachers may seem like a daunting undertaking. However, for many subjects and grades, the method described in this report relies on testing and other data collection that may already be going on (e.g., end-of-course testing and use of commercial tests that might be used as predictors). It does rely on common measures across groups of teachers, but as mentioned above, collaboration across teachers, schools, and even districts can have many efficiencies and can possibly result in better end-of-course measures, too. Further, data management systems are now commonplace in schools, providing a “home” for student test scores and grades. Given the emphasis on data use in our nation’s schools today, it should be rare that any school would not have personnel who are proficient at basic data analysis, and so implementing the regression analyses called for in this document should require someone to learn just a few more Excel commands. Should this capability be lacking in some smaller districts, then in collaborating with other districts to develop common assessments and obtain more teachers for the required analyses, perhaps districts could also consider cost sharing for analysis support. Most important, data gathering for purposes of evaluating teaching effectiveness need not be undertaken as a separate and distinct effort, but rather it should be accommodated by a district’s assessment system. If the approach recommended in this report is to be implemented, then for each subject-grade combination, a plan is needed that is consistent with the guidelines provided herein regarding end-of-course measures, predictors, numbers of teachers and students, etc. Once the data collection and analyses are implemented and routinized, they can lead to valuable information that can contribute annually to accurate and fair evaluations of teaching effectiveness and foster improved teaching and ultimately student learning.
Appendix A
Overview and Recommendations of the AASET Study

To inform the design of the desired system for using student growth in teacher evaluations, Measured Progress conducted a large-scale study involving pre- and post-testing with instruments in history, life science and mathematics that included both performance and multiple-choice components. Junior high and senior high teachers (and their students) from three states were involved. There were eighteen different subject-grade-state combinations for which separate analyses were completed and growth indicators computed.

For each subject at a grade in a state, there were multiple test forms randomly assigned to students. Every test included a multiple-choice component and a performance component, but only the latter component varied across forms. As an initial step in the analysis of student response data, the forms were statistically equated; therefore, they could be regarded as common or equivalent assessments. The same sets of forms were used in both pre- and post-testing, although the random assignment of forms for each administration meant that most students received different pre-test and post-test performance tasks. The fall and spring administrations were over six months apart. The scoring of performance components for both rounds of testing was completed after the spring administration by trained scorers at Measured Progress, not participating teachers.

The research team also collected prior academic performance data (previous year grades and state test scores) on the participating students. Final analyses focused on comparisons of results using different growth models and used data from over 12,000 students. Over 600 classrooms of students and approximately 250 teachers were involved. These teachers were the focus of effectiveness indicators computed using the various models.

Major Findings of the AASET Project

1. There were moderate to high correlations among results for different growth models applied to data from the same (or equated) tests, but anything less than “very high” means different models yield different results for many teachers.

2. Correlations between indicators based on the same model, but different end-of-year tests (state vs. AASET) or based on the same model applied to different test components (multiple-choice vs. performance) were quite variable (-.32 to .81).

3. Some analyses done separately with student scores subjected to sophisticated psychometric (Item Response Theory) scaling produced correlations among models that were only slightly higher than those based on raw (or linearly transformed) scores.

4. Teacher effectiveness indicators based on simple growth and simple prediction models correlated moderately to highly with indices from more sophisticated models (e.g., Hierarchical Value-Added Model [HVAM] and Student Growth Percentiles [SGP]).

5. For purposes of predicting post-test scores, the pre-tests were generally the best predictors, followed by prior-year state tests in the same subject, then prior-year state tests in other subjects. While previous course grades were the lowest correlates with the post-tests, they generally did tend to account for added variance in the post-test measures making them reasonable to use as predictors.
Recommendations from the AASET Project

1. Because most local districts do not have high-level psychometric expertise, the recommended approach for determining the extent of student growth in the non-tested subjects and grades should involve less sophisticated analyses than growth models such as the Hierarchical Value-Added Model (HVAM) and Student Growth Percentiles (SGP). The analyses should be readily conducted with commonly used software packages such as Excel.

2. Student growth statistics in a subject at a grade should be based on common assessments across teachers and involve enough teachers from within or across districts to generate meaningful normative data (i.e., data showing what might be average or expected student growth).

3. Different results for different tests or test components suggest that end-of-course tests should represent a good sampling of both the content and cognitive process levels (depth of knowledge) valued in the course. In the case of the latter, this means the inclusion of performance components is strongly advised. The lack of such a component would encourage an overemphasis on low-level knowledge in instruction.

4. Limited comparability of growth or value-added indicators across teachers leaves it to the teacher evaluator to apply human judgment in deciding if the student growth for a particular teacher is adequate given the unique characteristics of the teacher’s students, the other unique contextual factors of the teacher’s situation, and previous growth indicators for the teacher. (Evaluators’ capabilities to make such judgments is a whole other matter with implications for training, professional development, and supervisory monitoring.)

5. A simple prediction model for measuring student growth is recommended. Such a model would subtract students’ predicted scores on a common end-of-course measure from the actual scores on that measure, and aggregate (average) the differences at the teacher level. A simple prediction or regression model is recommended over simple growth (post-test minus pre-test) for several reasons, including availability of appropriate pre-tests and greater susceptibility of pre-post testing to “gaming the system” in the context of teacher evaluations.
Appendix B
Instructions for Performing Regression-Based Growth Analysis at the Teacher Level Using Excel

These instructions will reference screen shots from Excel worksheets associated with the application of the Excel software to real test data. In order to perform the multiple regression analysis, the Data Analysis Add-In must be activated in Excel. This is an optional component in Excel, but is included with the software. To activate this add-in for your version of Excel, please refer to your version’s documentation, or the Microsoft website: http://office.microsoft.com/en-us/excel-help/quick-start-activate-and-use-an-add-in-HA010370161.aspx.

Additionally, all instructions pertain to Excel 2013 for Windows. Actual location of tabs and buttons may vary on other versions of Excel. See your manual, or the Microsoft website for more information.

The directions below refer to a worked example of regression analysis using real data from students in eighth grade history classes taking common assessments.

**Step 1. Prepare student data to compute the regression.**

In order to compute the regression, the student data must be entered or imported into an Excel worksheet (or equivalent in another analysis system). In our sample worksheet, the raw data appears on the first tab we named "Soc08." The data required for completing these analyses are: teacher ID field, predictor scores, and end-of-course test score.

In our file, the teacher ID field is the variable labeled “Class” and is in column A; the predictors are labeled, “ReaScaledScore,” “SocNumberGrade,” and “rawscore_pre” and are in columns B though D; and the end-of-course test score is the students’ post-test score and is labeled “rawscore_post,” located in column E. A portion of the Excel worksheet is shown below:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Class</td>
<td>ReaScaledScore</td>
<td>SocNumberGrade</td>
<td>rawscore_pre</td>
<td>rawscore_post</td>
</tr>
<tr>
<td>2</td>
<td>Class</td>
<td>718</td>
<td>4</td>
<td>0.903099247</td>
<td>0.390501032</td>
</tr>
<tr>
<td>3</td>
<td>Teacher IDs</td>
<td>705</td>
<td>3.7</td>
<td>0.137597462</td>
<td>1.705141517</td>
</tr>
<tr>
<td>4</td>
<td>redacted</td>
<td>739</td>
<td>4</td>
<td>0.775515616</td>
<td>-0.68511391</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>700</td>
<td>4</td>
<td>0.903099247</td>
<td>0.629526575</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>632</td>
<td>4</td>
<td>1.158266509</td>
<td>0.629526575</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>718</td>
<td>4</td>
<td>2.051351924</td>
<td>1.466115974</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>706</td>
<td>4</td>
<td>1.143433777</td>
<td>0.749039346</td>
</tr>
</tbody>
</table>

There were actually several hundred lines in this worksheet, each line showing the data for an individual student. To put the data in context, some explanation of the variable names is in order. The teachers for whom student growth statistics were ultimately to be computed were teachers of 8th grade social studies (U.S. History). The data used in this analysis came from a study that involved pre- and post-testing. Thus, our end-of-course social studies measure is labeled rawscore_post. Our same-subject high quality predictor is the pre-test from the study, rawscore_pre. For purposes of the study which had students taking non-equivalent forms, the raw scores were converted to a common scale—hence, the values of these scores may not look like typical test scores. They are z-scores with a mean of 0.0 and standard deviation 1.0. (A z-score is just the number of standard deviations away from the mean that corresponds to a particular raw score. For example, a raw score of 65 on a test with a mean of 50 and standard deviation of 10 would correspond to a z-score of +1.5 since 65 is one and a half standard deviations above 50.) However, such transformations of scores have no effect on the analyses we are performing. SocNumberGrade is our same-subject previous year course grade. ReaScaledScore is actually the scaled score on a previous spring state reading test. That was our high quality test in another subject.
Step 2. Perform the regression.

Under the data tab, there is a “DataAnalysis” button. Clicking on this brings up a dialog box with several analysis tools. Select “Regression,” and click “OK.” This will bring up the dialog box for performing the regression. Fill in the ranges for your X and Y variables, and make sure the “Labels” and “Residuals” check boxes are checked. Selecting “New Worksheet Ply” will place the results into its own worksheet. See the screen shot below:

“Range” in Excel parlance above means the location of the raw data in the worksheet. In this case, the “Y Range” or the dependent variable (end-of-course test) range was rows 1 through 310 in column E. The predictor data was in rows 1 through 310 in columns B through D.
The results of this regression analysis can be seen in the sections of the worksheet shown below, which would be displayed under the second Excel tab that would be named “Soc08 Regression Results”:

\[
Y \text{ (pred)} = 0.006413 \text{ ReaScaledScore} + 0.226569 \text{ SocNumberGrade} + 0.357161 \text{ rawscore\_pre} - 4.90243
\]

The right-hand side of the equation is just the linear combination of predictors using the coefficients determined by the regression analysis and shown in the display above.
Step 3. Collate the residuals with the teacher IDs from the original student file.

The next step is to collate the students’ residuals with the teacher identifiers. To do this, a new sheet titled “Soc08 Collated” is created. Notice in the display below that the column with the teacher IDs from the “Soc08” tab has been copied and pasted into column A on this sheet, and the residuals were then copied and pasted into column B on this sheet from the Regression Results sheet. Of course, the IDs were redacted from the display in this document.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Class</td>
<td>Residuals</td>
</tr>
<tr>
<td>2</td>
<td>Teacher IDs</td>
<td>1.199113</td>
</tr>
<tr>
<td>3</td>
<td>Redacted</td>
<td>-1.70499</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-0.18582</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.159128</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.12523</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>-0.28705</td>
</tr>
</tbody>
</table>

It is important that the student data not be sorted or reordered between the calculation of the residuals, and copying and pasting of data into this table. The proper match between the students’ residuals and the students’ teachers is obviously critical.

Step 4. Compute teacher level results.

To compute the teacher level results, select the insert tab, then click the “Pivot Table” button. Then select the table and range that locate the input data needed, select “New Worksheet,” and click “OK.” See display below:

This will create a new worksheet, which would be labeled “Soc08 Teacher.” When this sheet first appears, you need to select all pivot fields, and then click on the “Values” popup list in the lower right, select “Value Field Settings,” select “Average,” and click “OK.” This will produce the following table:
This table shows the average student “growth” for each teacher—actually the average of the students’ differences between their actual and predicted end-of-course test scores. This is the teacher statistic the general approach is intended to produce. The teacher statistics can then be rank ordered, quartile groups could be determined, or any other secondary analysis could be performed. These statistics are expressed in the metric of the outcome or end-of-course measure. Recall, in this case, a student’s score was a z-score, which is the number of standard deviations from the mean the student scored on the test. Thus, referring to the table, the students of Teacher 14 on average scored 0.66 or two-thirds of a standard deviation above their predicted scores. If the raw scores on the test actually had a standard deviation of 15 points, say, then on average the teacher’s students scored 10 points above their predicted scores.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of Residuals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Tchr IDs Redacted</td>
<td>-0.279557309</td>
<td></td>
</tr>
<tr>
<td>Redacted 2</td>
<td>-0.040526274</td>
<td></td>
</tr>
<tr>
<td>Redacted 3</td>
<td>-0.331245595</td>
<td></td>
</tr>
<tr>
<td>Redacted 4</td>
<td>-0.022817305</td>
<td></td>
</tr>
<tr>
<td>Redacted 5</td>
<td>-0.345369019</td>
<td></td>
</tr>
<tr>
<td>Redacted 6</td>
<td>0.14744184</td>
<td></td>
</tr>
<tr>
<td>Redacted 7</td>
<td>-0.101493861</td>
<td></td>
</tr>
<tr>
<td>Redacted 8</td>
<td>0.139561525</td>
<td></td>
</tr>
<tr>
<td>Redacted 9</td>
<td>-0.072630028</td>
<td></td>
</tr>
<tr>
<td>Redacted 10</td>
<td>-0.172631302</td>
<td></td>
</tr>
<tr>
<td>Redacted 11</td>
<td>0.018566598</td>
<td></td>
</tr>
<tr>
<td>Redacted 12</td>
<td>0.798130715</td>
<td></td>
</tr>
<tr>
<td>Redacted 13</td>
<td>-0.79853801</td>
<td></td>
</tr>
<tr>
<td>Redacted 14</td>
<td>0.662552375</td>
<td></td>
</tr>
<tr>
<td>Redacted 15</td>
<td>0.739563734</td>
<td></td>
</tr>
<tr>
<td>Redacted 16</td>
<td>0.236274421</td>
<td></td>
</tr>
<tr>
<td>Redacted 17</td>
<td>0.40214997</td>
<td></td>
</tr>
<tr>
<td>Redacted 18</td>
<td>-0.149834148</td>
<td></td>
</tr>
<tr>
<td>Redacted 19</td>
<td>0.313726536</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>4.70947E-16</td>
<td></td>
</tr>
</tbody>
</table>