Student Growth in Teacher Evaluations:
Practical Issues in Designing a Fair, Accurate and Understandable System for Calculating Teacher Effects

Prepared by Bob Thompson, Education Consultant

All school districts in Alaska are currently in the process of developing measures that would comply with new state regulations concerning teacher evaluations. Of major concern to many districts is how to design a system that factors a student growth metric into teacher effectiveness evaluations. A few districts have worked on this issue as part of a requirement for federal grant money under School Improvement Grants administered in the last few years. The experiences of those districts, as well as the experience of many districts across the nation, provide some very practical advice and highlight some issues that every district that implements such a system must address.

This paper outlines some of those issues and offers a few suggestions on how to deal with them. Many problem-solving discussions will be necessary to determine which issues apply to a school, depending on the model(s) adopted, and how to solve the ones that do apply. This list is not comprehensive, but should serve to bring attention to the complexity of this task and the number of related issues associated with measuring teacher effects on student growth.

Student Rosters

The issue is which students will be used to quantitatively measure teacher effects on student growth for particular teachers. Problems arise when student placement changes during the school year, students are taught by multiple teachers, leveled grouping of students for instruction mixes students other than by grade level, and when determining instruction responsibilities for special education teachers when students are included in regular education classrooms.

Issues:
- Student management systems assign students to teachers according to subject and/or grade level. Instructional practice often mixes students in complex ways that are not reflected on teacher rosters.
- Tracking student movement and percentage of time spent receiving instruction on a particular subject from a particular teacher can be very complex and very time consuming.
- Student mobility from school to school and district to district can present issues with availability of usable data for the student and effects of prior learning on future teacher effects.

Suggestions:
- Consider these issues and write policies and procedures for handling the issues into teacher evaluation documents prior to attempting to implement a system. Determine how much and what data needs to be kept, and develop procedures for recording that data throughout the school year.

Mobility and Attendance

Student mobility and attendance can dramatically affect student performance. Issues when using student data for measuring teacher effects include what are the cutoff dates for enrollment and attendance, what attendance figures should be used with A.M. and P.M. attendance and multiple period attendance records, is there a maximum absence threshold where students will no longer be counted in student growth calculations, how do tardies affect growth calculations, is there differentiation between excused versus unexcused absences, does providing a student make-up work override an absence, and what is the cutoff date for students to be included when they transfer from another school or another district?

Issues:
- Attendance records are often poorly kept and unreliable resulting in large discrepancies between multiple periods and between A.M. and P.M. attendance. Example: A student has recorded absences in a six-period day of 41 – 28 – 6 – 4 – 10 – 18. If period 1 is Language Arts, what effect do the absences have on the students reading score? Is that score still valid for use in determining teacher effects?
- Elementary schedules often demand that subjects are taught at different times of day. If teacher A teaches math first thing in the morning and teacher B teaches math right after lunch, is there a system in place for using A.M. attendance for teacher A in math and P.M. attendance for teacher B in math?
- Considering each student's schedule and determining whether or not the student received instruction in a particular content subject measured by the chosen assessment can be very complex and time consuming. Example: High school student A has two math classes, algebra and remedial math, taught by two different teachers. How are teacher effects on the summative math score determined?
- Many high schools deny credit for students with twelve or more absences in a class during a semester. Should these students be used for calculating student growth if all other data is available?

Suggestions:
- Establish clear guidelines for handling these and other issues regarding attendance. Multiple period discrepancies can be particularly difficult. A system must be devised that can calculate growth based on guidelines without undue recordkeeping and complex calculations.
- Until studies that provide more-definitive guidance become available, systems might be advised to assign teachers responsibility only for students who spend most of the year with them.

Data Integrity

Regulations require that data used to determine teacher effects on student growth be objective, empirical and valid. Across the nation many systems have been adopted that use varying value-added measures (VAM) to calculate teacher effects. Some states have developed very complex VAMs to be used statewide. In Alaska, the task is left up to individual districts to devise systems that use objective, empirical and valid measures to measure teacher affects on student growth. The selection of these measures requires teachers' involvement, comparisons of measurements before the student is taught and after the student is taught, and that the knowledge, understanding, or skill is taught during that student's time with the teacher.

Issues:

- State regulations require that data from summative statewide tests must be used for teachers if "…the data is directly related to the job duties of the educator." However, a minimum of two measures is required for each educator. For teachers that do not have job duties directly related to the data from the summative assessments, two additional measures apart from the state assessments must be found.
- In addition to state assessments, many districts use either MAP or AIMSweb as an additional measure. Neither of these assessments was designed as a VAM and particular problems occur when trying to use these assessments above the eighth grade.
  - Using AIMSweb targets to measure teacher effectiveness measures student performance against a benchmark, such as LNF, PSF, R-CBM, or M-CAP targets. Students can make considerable growth and not meet the benchmark by the end of the year. To measure teacher effectiveness the Rate of Improvement must be used, or districts will need to establish growth ranges for students based on pre- and post- tests to assign evaluation performance levels.
  - AIMSweb norms tables for R-CBMs only go through the 8th grade. Data shown on tables for grades nine through twelve are the 8th grade data. Using these tables to measure high school student growth is not valid.
  - MAP growth projections are based on national norms for students at a specific grade level and a particular RIT score made in the fall. These norms only go up to the 11th grade. Norms for 12th grade students are not provided by NWEA and most N sizes for 11th grade students and some 10th grade students are too small to provide growth projections.
  - Norms for MAP growth get smaller as students get older so that nearly all growth projections for students in 9th grade and above

are between 2 and 3 RIT points. In reading the growth projections from fall to spring are less than 1 RIT point at these grade levels. This growth projection is typically less than the standard error of measurement for a student taking a MAP test making the growth projections invalid.

   o Using the standard error of measurement, NWEA MAP RIT scores provide a range for which the assessment scores are valid. The specific RIT score used to calculate growth is in the middle of that range. Should the high end of that range be used to calculate actual growth in order to make sure the calculation does not rely on a score that is too low?

- Other screening assessments, such as STAR, DIBELS, and easy CBM present the same problems as AIMSweb early literacy screening tools.
- Case studies by IES and the Center for American Progress show that using multiple measures for teacher effectiveness produces more valid and reliable results than using one or two measures. These means districts must find even more measures of teacher effectiveness than are commonly used in Alaska districts today.
- In order to comply with requirements that teachers instruct students in skills and knowledge that are measured for teacher effectiveness many states and districts have developed their own assessments to measure content aligned to standards and particular content; however, ensuring reliability and validity of student growth measures requires expert analysis of the methodology that districts usually do not have.
- Using aggregate student performance measures to evaluate teachers in non-tested subjects or grades allows school systems to rely on existing measures but creates a two-tiered system in which some teachers are evaluated differently from others.

Suggestions:
- Use as many measures of student performance as possible to increase the reliability of measuring teacher effectiveness.
- Use commercially available or state approved measures whenever possible so that reliability and validity are already established. It is important to ensure that the content measured by these assessments is consistent with the content teachers are being asked to teach.
- Use multiple years of student achievement data in value-added measurements, and, where possible, average teachers' value-added measurements across multiple years.
- Develop a MAP Index that uses a variety of MAP data measures from a single test administration, such as meeting growth projections, ratio of students showing growth, mean RIT growth, and average growth, to form a single number to be used to measure growth, instead of relying solely on the system generated growth projections.

- Develop content specific assessments from program assessments and teacher-developed assessments for use with secondary content teachers.

System Integrity

Students not tested and without appropriate data present problems when deciding which students to include in teacher effectiveness measures and how to include those students in calculations. Inaccuracies in keeping attendance and failure to record changes in student placement can create situations where the validity of the data is questioned. Consistent test administration becomes paramount when trying to compare growth of one group of students with another group of students. Scores from state assessments may not be available in time to affect evaluations and approval of contracts for the next school year. Incongruency between data collection systems, such as student management systems and data collection systems, can create problems in aligning data for specific purposes.

Issues:
- Should alternate systems be devised to create data for students without scores or growth projections, such as the NWEA Growth Calculator for creating growth projections?
- When student scores decline significantly in a progressive scaled measure, such as a RIT score or Lexile score, from one measure to the next, should these students be re-tested, not used in calculations, or used, although the measure reflects an inaccurate measure of student knowledge and skill?
- Changes in assessments means projections based on previous tests may not be valid when compared to new tests.
- When students are excluded from measures of teacher effectiveness, there is a reduced incentive for the teacher to affect that student's performance.
- Tenured teacher contracts must be issued by March 15th of each school year, which is prior to summative state testing of students.

Suggestions:
- Ensure that procedures are in place for makeup testing to make certain that all enrolled students have data to be used for teacher effectiveness measures.
- Increase accountability for teachers to keep accurate attendance data and to record any changes in student placement.
- Provide professional development to ensure inter-rater and test administration reliability for test administrators and scorers.
- Consider averaging teacher measurements over multiple years. Case studies show that this increases measurement reliability.

- Use the prior year's state test data for teacher evaluations. First year teachers would have to be exempted from using this data.
- Align data systems so that data can be easily matched to students and teachers for student growth ratings.
- Consider development of a database to be used exclusively to hold all relevant data to be used in evaluating teacher effectiveness.
- Devise measures to hold teachers accountable for improving student performance for students not used in teacher evaluations.

Conclusion

Quantitatively measuring teacher effects on student growth is a complicated endeavor that must attempt to isolate teacher effects from non-school related factors. At the same time, the system must account for numerous complexities involved in the daily practice of instructing students in a dynamic environment.

Further complicating things is the difficulty communicating complex methodologies and concepts to teachers, parents, students and communities that have an interest in the outcomes. For the system to work, clarity, transparency, stakeholder involvement in development, and psychometric validation of methods are essentials.

As states and districts across the country are realizing, being on the leading edge of change means there will be trial and error, mistakes and successes, lessons learned and new ideas tested, and progress will be made in fits and starts.

Is it worth the effort? Time will tell. Research is being done and preliminary studies are showing that VAM systems and valid measurements can produce reliable results. Measurements being used across the country are having a high correlation to actual student learning. Case studies are currently being completed with new data available in the spring of 2014 from IES.

To be able to accurately evaluate teacher effectiveness means to be able to define what it is that teachers do that makes student learning possible. Teaching is an extremely complex task and quantitative measures will never reach the point of totally defining the behaviors that make for great teaching; however, quantitative measures are what we have to determine whether or not students are learning. Using those measures to determine an important aspect of teacher effectiveness is only sensible.

If you have further questions or need clarification on any of these issues, please feel free to contact Bob Thompson, Expect Educational Excellence, in Palmer, Alaska, at (907) 745-2019 or littlesusitna@hotmail.com

Bob Thompson is an education consultant with 28 years of experience in teaching and public school administration in Alaska. He has taught in the Mat-Su School District and has been an administrator in Mat-Su, North Slope and Southwest Region school districts. He has owned his own education consulting business since 2006 and has served on state committees, conducted curriculum audits, and contracted as a State System of Support School Improvement Coach. He currently serves as co-curriculum director for the Yupiit School District and is the SSOS Coach for Henry Lott Memorial Elementary School and Tuluksak High School in Tuluksak, Alaska.