

Using Student Growth Data in Teacher Evaluation: Selecting Measures for the Other Teachers' Students

Cheryl A. Gould

In an effort to ensure schools have highly effective teachers in the classroom, Virginia and other states developed teacher evaluation systems that include student growth data as a substantial portion of teacher evaluation. Some teachers, mainly reading and math teachers, have access to student growth data based on recently developed value-added estimates or Student Growth Percentiles calculated from standardized assessments.

Yet, approximately 69 percent of K-12 teachers do not teach reading or math and therefore have no such available measures. Nevertheless, most states require these teachers to produce evidence of student growth just like the reading and math teachers who have available value-added data (Prince, 2009). Many of the 69 percent struggle to identify valid and reliable student growth measures for courses where no value-added measures or standardized assessments exist. When it comes to assessment of student growth, the stakes are higher now than in the past because administrators use assessment results not only for instructional decision-making but also for human resources decision-making concerning teachers' jobs. Assessment results must hold up to the rigorous scrutiny from administrators during the teacher evaluation process. In addition, teachers must accept the selected measures as a fair representation of student growth and teaching contributions.

This article provides a thorough discussion of (a) policies and practices leading to the use of student growth data in teacher evaluation, (b) a brief discussion on widely accepted student growth models, and (c) alternative assessments (non-standardized) that might be used to assess student achievement and growth, with a focus on authentic assessment as a viable measure of student growth.

Policies and Practices Leading to the Use of Student Growth Data in Teacher Evaluation

President George W. Bush signed the No Child Left Behind Act (NCLB) of 2001 into law in an attempt to close the "achievement gap" through increased student learning, teacher quality, and school responsibility by way of test-based accountability and other requirements (Baker et al, 2010). In an effort to increase teacher quality, NCLB requires that teachers in core subjects be "highly qualified" (Silva-Mangiante, 2011). In order to be highly qualified, a teacher must hold a bachelor's degree in the teaching subject, have full state licensure, and demonstrate subject matter competence (US Department of Education, 2004). This requirement for highly qualified teachers ensures teachers in the classroom have the prerequisite skills needed for effective teaching.

More recently, the American Recovery and Reinstatement Act of 2009 called for states to improve teacher effectiveness and "address the inequalities of teacher distribution of highly qualified teachers" (H. R. 1—169). In many states, higher quality teachers tend to be more concentrated in affluent schools while lower quality teachers tend to be more concentrated in the least affluent schools (Gordon, Kane, & Stringer, 2006). Competitive funds provided in the Act may go to schools that produce *evidence* of increased recruitment and placement of high-quality teachers in low-performing schools, as "demonstrated with meaningful data" (H.R. 1-170). This federal initiative, referred to as Race to the Top, requires schools to use multiple measures to gauge teacher effectiveness with a substantial portion coming from student growth data. School divisions may use teacher evaluation results "to inform human capital decisions such as: professional development, compensation, promotion, retention, tenure, and removal" (US DOE, 2010, p. 34 as cited in Silva-Mangiante, 2011).

States have moved quickly to meet the Race to the Top requirements by increasingly using student achievement data with a focus on improving teacher effectiveness. For example, New Jersey's policies require schools to base 35 percent to 50 percent of a teachers' evaluation on student growth data. Teachers who receive two consecutive years of ineffective or partially effective ratings are subject to personnel actions. Another school system, the District of Columbia (D.C.) Public Schools, identified two separate teacher groups in their teacher evaluation system entitled *Impact* (D.C. Public Schools, 2013; Silva-Mangiante, 2011). English and math teachers in the D.C. district have a different evaluation format than all other teachers in the district. The D.C. district justifies this bifurcated system with the fact that there are available statistical methods for calculating long-term student growth in reading and math but not for other subjects.

In Virginia, the state's Department of Education (2011) requires school divisions to base 40 percent of the teacher's evaluation on student growth data. In order to connect student performance to teacher evaluation, the VDOE recommends "20 percent of the teacher evaluation (half of the student academic progress measure) come from Student Growth Percentiles as provided from the Virginia Department of Education when the data are available and can be use appropriately" (p. 42) and the other 20 percent from "one or more alternative measure with evidence that the alternative measure is valid" (p. 42). In those cases where there are no readily available validated measures, the VDOE encourages teachers and schools to implement *student goal setting* as a way of measuring student progress.

New Jersey, D.C., and Virginia are three examples of how states are handling the federal requirement that states base a substantial portion of teacher evaluation on student growth data. So, *why* are federal and state policy makers demanding schools use student growth data in teacher evaluation? After all, school divisions have been documenting student proficiency in academic core courses for a decade or more. To understand this issue, one must become familiar with the recent shift in K-12 education from measuring student

proficiency to measuring student growth, educational developments in the area of student growth data, and research findings using student growth data.

Measuring student proficiency is different than measuring student growth. To measure student proficiency, one looks at how the student performed on a given set of standards or criteria at any given point in time as measured by a test, quiz, project, etc. Whereas, measuring student growth is the ability to track student performance over time by comparing proficiency at the beginning of an instructional period and at the end of the instructional period. Division or state educational leaders can track student growth over months, years, or several years using a variety of methods (Sanders & Rivers, 1996; Sherrer, 2011; Stronge, 2010). Two popular methods for measuring student growth are Student Growth Percentiles and value-added estimates.

Both Student Growth Percentiles and value-added estimates measure student growth and provide data for decision-making, but value-added estimates have a distinct difference. Value-added methods can estimate the "value-added" from the teacher's instruction by controlling for variables not attributed to the teacher such as school effects or socio-economic effects on student learning. While there is strong support for using such models in assessing teacher effectiveness (Sanders, 2003; Stronge, 2010), it is important to note that there is ongoing scholarly debate on the reliability and validity of the results produced by these systems (Herman, Heritage, & Goldschmidt, 2011). Nevertheless, research using these models has yielded unprecedented research-based findings on teacher effectiveness.

For example, in one study involving value-added methods, Sanders and Rivers (1996) revealed that the quality of teacher instruction has both a cumulative and residual effect on student achievement. They found that the most effective teachers facilitated learning for all students while the least effective teachers had unsatisfactory gains for all students. As teacher effectiveness increased, low achieving students were the first to make gains followed by average students. They also found that cumulative effects on learning were substantial. A

student with a high quality teacher over several years continued upward achievement. On the other hand, if a student had a low quality teacher for several years the impact on achievement was severe. These findings underscore the critical importance of having highly effective teachers in every classroom in the nation so that all students have the opportunity to grow.

In short, inadequacies in policy mandates such as NCLB (2001) prompted educational researchers to develop value-added and Student Growth Percentile models. These models allowed educational researchers to gather data leading to significant findings on teacher effectiveness. Unprecedented findings from such studies informed legislators to enact recent policy mandates involving the use of student growth data in teacher evaluation. This cycle involving policy, educational development, and educational research is likely to continue. Until there are further educational developments and research in the areas of alternate assessment used for measuring student growth, teachers are left to identify viable alternative measures from those that already exist or to develop new measures at the local level.

Selecting Assessments for Measuring Student Growth

Student Growth Percentiles and value-added methods are useful tools for estimating teacher effectiveness in specific subjects where there is clear vertical curricular alignment and available state standardized tests, such as reading, math, and sometimes science. Yet, as mentioned previously, 69 percent of K-12 educators teach courses that have no associated standardized assessments, no vertical curricular alignment, or only loosely aligned curricula with some content overlap. Special Education or English Language Learner teachers may have classes too small to gather statistical information. For these reasons, 69 percent of teachers cannot use value-added or Student Growth Percentile models (Prince, 2009). These teachers must ask, "What assessment(s) will I use to provide reliable and valid evidence of student growth over one year's timespan or perhaps a portion of that year?" In order for educators to

select the most accurate measures, one must understand the variety of assessments options.

Educational professionals use the term *assessment* in many contexts. For example, there are formative and summative assessments, standardized assessments, performance assessments, classroom-based assessments, and on-the-spot assessments. Assessment in the school setting typically falls along a continuum of traditional to authentic. Oloruntegbe (2010) provides a description of the continuum.

Traditional \longleftrightarrow Authentic

Events at the two ends of the continuum range from *selection of response to performing a task*; from *contrived to real life*, from *recall or recognition to construction and application*. Four categories rest within the continuum: (a) tests, (b) product or project assessment, (c) performance assessment, and (d) process skills assessment (Oloruntegbe, 2010).

There is no clear consensus of which assessments are best but no matter the type of assessment used, the practical value of the assessment is most important (Guskey, 2007; Tomlinson, 2001). That is to say, we must use assessment to determine the extent to which a student has developed targeted knowledge and skills. Assessment scores are not an end to themselves; rather they are *measures* of achievement and growth used to determine a course of instructional action for a student or a group of students (Popham, 2009).

Authentic Assessment- A Viable Student Growth Measure

Authentic assessment provides a viable measure of achievement and growth. Scholarly literature supports authentic assessment as a reliable and valid measure that can withstand scrutiny. The term authentic assessment came about because in this type of evaluation, educators ask students to perform *real world* tasks and therefore, consider the assessment more authentic than traditional forms of assessment (Darling-Hammond, 1994; Oloruntegbe, 2010; Svinicki, 2004; Tanner, 2001). Authentic

assessment may be of particular interest to those who teach courses in career and technical education, art education, music education, or other disciplines with an emphasis on real-world application.

Authentic learning activities come in many forms such as projects, products, portfolios, or performance tasks, to name a few. No matter the form, authentic assessments share common characteristics (Dana & Tippins, 1993; Svinick, 2004).

- Content focuses on big ideas, not insignificant facts.
- The student must use problem solving, further inquiry, judgment, and innovation.
- Efforts may result in a quality product or performance, rather than a correct response.
- Students can display strengths because the focus is on what the student knows, demonstrated by individual viewpoints, interpretations, and presentation/demonstration styles.
- Teacher and student establish and understand task criteria before the task begins.
- Scoring, often using a rubric, focuses on the quality of the task rather than ease of scoring.
- Diagnostic information and feedback may lead to immediate student improvements.
- Students complete tasks safely and easily in a classroom while preparing student for “outside” situations.

These characteristics may make authentic assessment a better predictor of student performance beyond school as compared to other assessments. “Because authentic assessments track the real world so closely, they are likely to have a great deal of face validity both for students and for any outside evaluator (Svinicki, 2004, p. 27) To withstand scrutiny, authentic assessment should provide (a) criterion-based standards, (b) multiple indicators of quality, and (c) some provision for judgment reliability (Tanner, 2001).

Issues surrounding reliability in authentic assessment focus on reliability in grading. Yet, evaluators achieve reliability in grading in different ways in authentic assessment than in traditional assessment. The use of objective, dichotomously scored test items, that eliminate the need to

determine the *degree* of correctness, improves reliability in traditional testing (Tanner, 2001). Controlling testing circumstances through “sameness” also increases reliability. For example, students answers the same questions or same types of questions, have the same amount of time, and use the same testing implements (e.g. paper, #2 pencils, one piece of scratch paper provided by the test administrator). These efforts increase reliability of results in traditional assessment.

In authentic assessment, teachers expect conditions to vary. The ability to adapt related instructional activities to the diverse needs of learners is one benefit of using authentic assessment. In fact, authentic assessment takes into account diverse backgrounds, skills, and abilities of students, ultimately “requiring that each learner, construct, what is for them, an original response” (Tanner, 2001, p. 26). Authentic assessments are more sensitive to differences among learners than traditional assessments and therefore are a more valid measure of student learning (Moon, Brighton, & Callahan, 2005). However, this lack of “sameness” may result in problems with scoring reliability.

Authentic assessment requires educators to judge multiple unique finished products based on differing student choices, understandings, and diverse backgrounds. This uniqueness of product represents the essence of authentic assessment. Even so, educators must find a way to maintain scoring consistency from student to student.

Educators may achieve scoring consistency in two ways. First, scoring instruments such as rubrics allow evaluators to assess identified assignment criteria by varying degrees of quality while maintaining consistency from student to student. In some instances, instrument developers categorize degrees of quality using qualitative terms such as “novice, competent, proficient, or expert” or scored in quantifiable terms such as “1, 2, 3, or 4”. “Ideally, authentic assessment employs detailed verbal descriptions of performance rather than the semi-cryptic letter grades or scores” (Tanner, 2001, p. 26).

Second, the use of multiple raters adds an element of fairness and constancy to evaluations. Training evaluators on the appropriate use of

evaluation instruments can improve inter-rater reliability. "Developing a reasonable level of inter-rater reliability is difficult and expensive because of the training required. The alternative, however, is to allow the instructor's judgments be a major source of assessment error" (Tanner, 2001, p. 28).

Authentic assessment validity has a different construct than traditional assessment validity. Discussion on traditional assessment validity may focus on content validity, criterion related validity, and construct validity. On the other hand, authentic assessment validity relates more to *assessment relevance* and *consequential validity* (Darling-Hammond, Aness, & Falk, 1995). Assessment relevance is the extent to which classroom learning relates to what the student must know and be able to do once leaving the classroom. The assessment results are valid if they predict how closely the student's performance on the assigned task conforms to what he or she will be expected to do later (Tanner, 2001). Darling et al. (1995) define consequential relevance as "the consequence for students and for schools of using a particular form of assessment" (p. 64). Validity lies in the relationship between teaching and learning. Put simply, if the particular assessment improves teaching and learning, then educators may consider it valid. If it does not improve teaching and learning, then the assessment lacks consequential validity. "Performance measures have the potential for increased validity because the performance tasks are themselves demonstrations of important learning goals rather than indirect indicators of achievement" (Resnick & Resnick, 1992).

As mentioned previously, advocates of authentic assessment proclaim that one benefit is the ability to level the playing field for students disadvantaged by culture, ability, or language (Allison & Rehm, 2006; Guskey, 2007, Tanner, 2001). However, authentic assessment can be problematic for students still learning English. Authentic assignments are language rich activities that often require student to make a presentation or explain a position or process. In this regard, authentic assessments may be more complex than multiple-choice and short-answer tests, putting English language learners at a disadvantage (Tanner, 2001).

Teachers as Stakeholders

Teachers are the primary stakeholders when using student growth data in teacher evaluation. Stakeholders' perceptions of the validity of a particular measure are an important factor, especially for those who will reap the consequences of accountability (Guskey, 2007). Lane, Parke and Stone (1998) explain that teacher and administrator perceptions of the meaningfulness and relevance of assessment results affect the motivation and effort they put forth to improve instruction and student learning outcomes. If teachers do not believe in the validity of a measure, then they will be less likely to support the effort and see it as a fair assessment of their teaching ability. Teachers may also need to educate and garner the support of colleagues and administrators who may be more familiar and therefore more comfortable with traditional forms of instruction and assessment (Ward & Lee, 2002).

Shepard (1995) warns, "Even authentic measures are corruptible and, when practiced for, can distort curriculum and undermine professional autonomy" (p. 38). The pressures teachers are now under to produce evidence of student growth may provide added incentive to skew evaluations results. Multiple raters and random checks of scoring practices guard against such reprehensible practices. "Assessment and grading continue to be a private activity, with considerable variation among teachers" (McMillan & Workman, 1998, p. 29). Using alternate assessment in teacher evaluation may bring this private activity into a more public light.

Given the vast literature supporting authentic assessment as a reliable and valid student performance and growth measurement tool, teachers should be encouraged to use authentic assessment. Yet, researchers suggest that the frequency with which teachers have used authentic assessment in the past varies for many reasons (Moon et al, 2005, Ohlsen, 2007).

First, teachers who are reluctant to use authentic assessment may not espouse the constructivist theory behind such measures (Ohlsen, 2007). Constructivist theory is the basis of most alternative assessments, including authentic assessment, in that learners construct their own

knowledge and make meaning based on prior experiences as well as past and current social interactions. One study cited by Ohlsen revealed that three groups of secondary teachers, social studies, foreign language, and math, were less constructivist than others were and therefore might be less inclined to implement alternate assessments in the classroom.

Next, teachers may lack confidence in using performance-based assessment (Ohlsen, 2007). Using authentic assessment is a new approach for many teachers and thus requires teachers to rethink the classroom assessment environment (Moon et al., 2005). Teachers need professional development, time, and a supportive social network to gain skills and learn strategies in using alternate assessment, especially authentic assessments such as those that are performance-based (Hattie, 2009; Rogers, 2003; Ohlsen, 2007). “Teachers will not just move from not doing a new behavior to doing it: they go through decision phases” (Hattie, 2009, p. 257). Rogers (2003) contended that adoption of any new practice is a process and teachers typically go through the phases of awareness, knowledge, persuasion, decision, implementation, and confirmation. These phases may overlap or may not occur for each person but the general idea is that supportive social networks are a powerful force in the adoption of innovative practices. Conversely, the lack of a supportive social network may become the biggest hurdle to innovation. Leaders may want to enlist the help of veteran teachers as mentors because experienced teachers are more likely to implement alternative assessments in the classroom than are their inexperienced colleagues (Moon et al., 2005).

Finally, alternative assessment is more time consuming to create, implement, and score than are test and quizzes (Ohlsen, 2007; Svinicki, 2004; Tanner, 2001). These assessments typically take (a) more planning on the part of teacher and student, (b) take up more class time to implement, and (c) take more time for assessment and reporting of results compared to results from machine-scored, multiple-choice tests. Guskey (2007) acknowledges the additional time involved but believes the trouble is worth it. Using a variety of measures will, no doubt, make reporting more complicated but failing

to do so restricts students’ capabilities and learning adventures.

Conclusion

There is ample research underlining the importance of having effective teachers in the classroom. Educational researchers have thoroughly documented the connection between quality instruction and student achievement. Teacher effects are substantial and long lasting. The development and implementation of student growth models such as value-added and Student Growth Percentile methods made such research possible.

There is scholarly debate on the reliability and validity of student growth models. The use of such models in conjunction with teacher evaluation has fueled the debate in recent years. Now, school divisions not only use student growth models to guide instructional improvements but also use them to assess teacher effectiveness and guide human resources decision-making. Given this new application, the use of student growth models has become a hot topic in education.

The majority of literature about student growth data in teacher evaluation centers on student growth models for reading and math because these are subject areas with vertically aligned curriculum from one year to the next. However, there is much less literature on how other teachers will provide student growth data for teacher evaluation. Educators must use other forms of assessment to fill the gap where no standardized tests and/or student growth models exist. District and state administrators, who tend to trust standardized tests over other forms of assessment, may question results gathered from alternate assessments.

Teachers, as primary stakeholders in this venture, are increasingly responsible for identifying alternate assessments and providing evidence of the validity of selected measures. How will teachers show that these alternate assessments provide valid results? How will district administrators, in turn provide evidence to state administrators that these alternate assessments provide valid results? State and division leaders have yet to answer these questions.

Literature on authentic assessment is abundant and there appears to be little challenge to

the use of authentic assessment as a viable measure of student performance. The literature provides thorough information on the characteristics, benefits, and challenges as well as constructive discussion on the reliability and validity of the authentic assessment. On the other hand, using authentic assessment to document student growth in the context of teacher evaluation is new research territory and therefore represents a topic about which educational researchers have written little. This is an area ripe for future research.

References

- Allison, B., & Rehm, M. (2006). Meeting the needs of culturally diverse learners in family and consumer sciences middle school classrooms. *Journal of Family and Consumer Sciences Education, 24*(1), 50-62.
- American Recovery and Reinvestment Act of 2009. Public Law H.R.1-111U.S.C. (2009).
- Anderson, S., Medrich, E., & Fowler, D. (2007). Which achievement gap? *Phi Delta Kappan, 88*(7), 547-550.
- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Ravitch, D., . . . Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. (Brief No. 278). Washington, D.C.: Economic Policy Institute.
- Danna, T. M., & Tippins, D. J. (1993). Considering alternative assessment for middle level learners. *Middle School Journal, 25*(2), 3-5.
- Darling-Hammond, L. (1994). *Setting standards for students: The case for authentic assessment*. New York: Teachers College Press.
- Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. Columbia University: New York: National Center for Restructuring Education, Schools, and Teaching, Teachers College.
- District of Columbia Public Schools. (2013). Impact. Retrieved June 15, 2013, from <http://www.dc.gov/DCPS/impact>
- Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job*. Washington, D.C.: Brookings Institute.
- Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice, 26*(1), 19-27.
- Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.
- Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version)*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement Issues and Practice, 17*(2), 24-28.
- McMillan, J. H., & Workman, D. J. (1998). *Classroom assessment and grading practices: A review of the literature*. Richmond, VA (Eric Document Reproduction Service No. ED453263): Metropolitan Educational Research Consortium.
- Moon, T. R., Brighton, C. M., & Callahan, C. M. (2005). Development of authentic assessment for the middle school classroom. *The Journal of Secondary Gifted Education, 16*(2/3), 119-133.
- No Child Left Behind Act of 2001, 20U.S.C. 6301 (2002).
- Ohlsen, M. T. (2007). Classroom assessment practices of secondary school members of NCTM. *American Secondary Education, 36*(1), 4-14.
- Oloruntegbe, K. O. (2010). Approaches to the assessment of science process skills: A

- reconceptualist view and option. *Journal of College Teaching & Learning*, 7(6), 11-18.
- Popham, J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, 48, 4-11.
- Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, DC: Center for Educator Compensation Reform. Retrieved from <http://www.cecr.ed.gov/guides/other69Percent.pdf>
- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford, & M. O'Connor (Eds.), *Changing assessments; alternative views of aptitude, achievement, and instruction* (pp. 37-75)
- Rogers, E.M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press
- Sanders, W. L. (2003). *Beyond No Child Left Behind*. (White paper No. 48.024). Presented at the 2003 Annual Meeting American Educational Research Association, Chicago, IL: SAS.
- Sanders, W. L., & Rivers, J. C. (1996, November). *Cumulative and residual effects of teachers on future student academic achievement*. (). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.
- Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin*, 95(2), 122-140.
- Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership*, 54(5), 38-43.
- Silva-Mangiante, E. M. (2011). Teachers matter: Measures of teacher effectiveness in low-income minority schools. *Educational Assessment, Evaluation & Accountability*, 23(1), 41-63.
- Stronge, J. H. (2010). *Effective teachers = student achievement: What the research says*. Larchmont, New York: Eye on Education.
- Svinicki, M. D. (2004). Authentic assessment: Testing in reality. *New Directions for Teaching and Learning*, 100(Winter), 23-29.
- Tanner, D. (2001). Authentic assessment: A solution, or part of the problem? *The High School Journal*, 85(1), 24-29.
- Tomlinson, C. A. (2001). *How to differentiate instruction in mixed-ability classrooms*. (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
- U.S. Department of Education. (2004). *New No Child Left Behind flexibility: Highly qualified teachers*. (Fact Sheet). Washington D.C.: Author.
- Virginia Department of Education. (2011a). *Guidelines for uniform performance standards and evaluation criteria for teachers*. Richmond, VA: Author.
- Ward, J. D., & Lee, C. L. (2002). A review of problem-based learning. *Journal of Family and Consumer Sciences Education*, 20(1), 16-26.