PACER RESEARCH for ACTION

ISSUE BRIEF
TEACHER EFFECTIVENESS:
An Update on Pennsylvania's Teacher Evaluation System

December 2013

Pennsylvania Clearinghouse for Education Research (PACER)

Introduction

Act 82 of 2012 established new standards for Pennsylvania's teacher evaluation system, including the incorporation of student performance measures in ratings decisions. Since 2009, approximately 35 states have amended teacher evaluation systems, with student achievement playing an increasingly prominent role.[i] This count includes neighboring states—such as New Jersey, New York, and Ohio—which base between 40 and 50 percent of teacher effectiveness ratings on student achievement.[ii,iii]

Changes to teacher evaluation policy have been motivated in large part by U.S. Department of Education's priorities, including the issuance of waivers from certain *No Child Left Behind* requirements. States receiving a waiver, including Pennsylvania, are required to "develop and implement teacher and principal evaluation and support systems that include student achievement growth as a factor."[iv] The nonpartisan Center on Education Policy reported that ten states amended their plans for a new teacher evaluation system due to the waiver policy alone.[v] Likewise, the Race to the Top (RTTT)[1] competition emphasized teacher and principal evaluation systems based on student achievement.[vi]

Proponents of these policies argue that linking measures of student achievement to determinations of teacher effectiveness provides educators with valuable feedback on their practice, and offers a means to recognize and reward teachers for their contributions to student learning. Opponents counter that certain measures do not provide valid or reliable indications of effectiveness as these measures don't take into account differences in non-school-based factors such as differences in participation in after-school programs, summer learning opportunities, and home learning environments among others. [vii]. Additionally, basing measures of teacher effectiveness and personnel decisions on standardized test results can produce unintended consequences such as a narrow curricular focus on tested subjects like reading and math at the expense of non-tested content.[viii]

This policy brief provides a closer look at Pennsylvania's new teacher evaluation system and the efforts of the Pittsburgh Public Schools—the state's second-largest district and an early adopter of revised evaluation standards—to implement reforms. We conclude with implications for state policymakers, district leaders, and education stakeholders.

[1] In Philadelphia, an $11 million RTTT grant will help launch the new evaluation system.

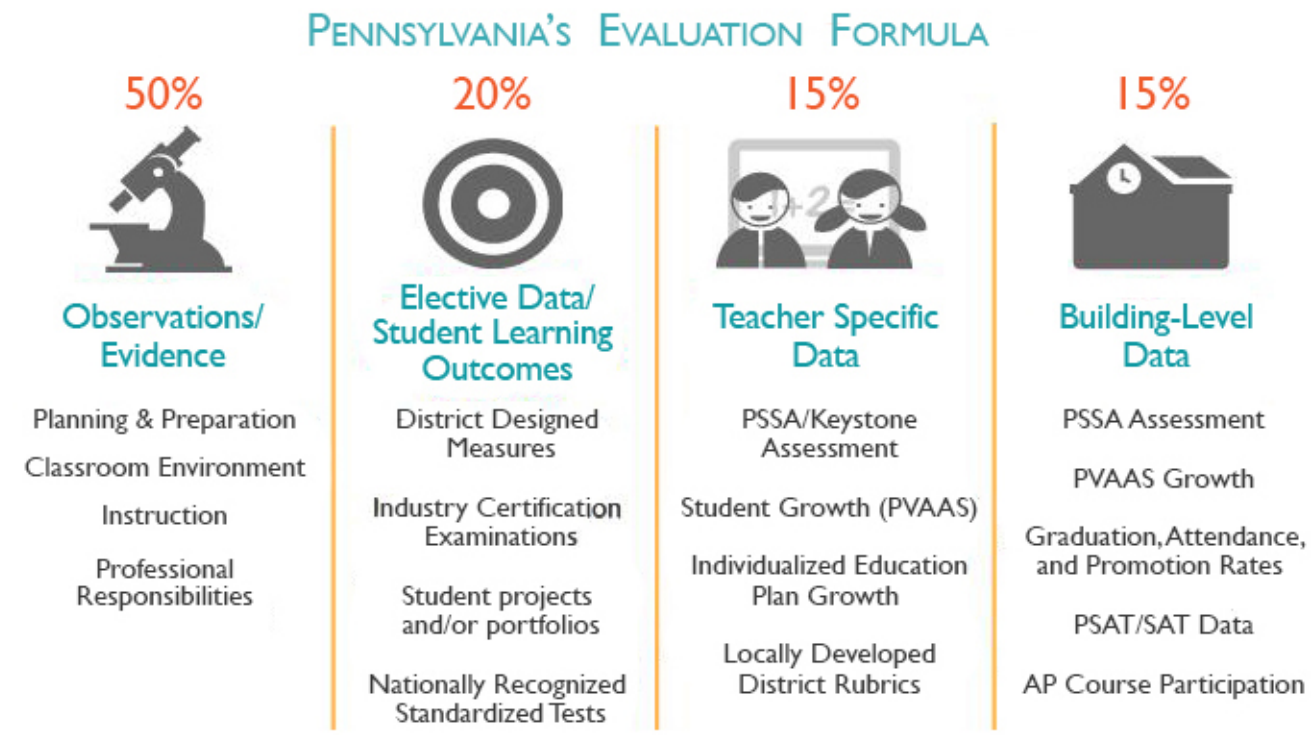
Pennsylvania's Evaluation Formula

Historically, Pennsylvania's system of teacher evaluation relied exclusively on classroom observations by school administrators or other staff, and was based on the Danielson Framework, which measures a teacher's planning and preparation, classroom environment, instructional strategies, and professional responsibilities.[ix] Research on districts nationwide that employ the same observation framework found "positive relationships between teacher evaluation scores and student achievement." The correlations between classroom observation ratings and student achievement were strongest when there were multiple, well-trained observers.[x]

Leading up to the passage of Act 82, Pennsylvania engaged in a three-year pilot program with districts throughout the state while hosting forums for teachers, administrators, and education leaders to discuss alternative approaches to evaluation. The first stage of the pilot involved four school systems, including a major urban school district and an intermediate unit, and analysis of statewide value-added data and its relationship with classroom observation results from the pilot sites. The second pilot phase expanded to approximately 100 districts statewide, and involved more than 5,000 educators.

A major theme emanating from this work was the need for multiple measures in evaluating teacher effectiveness. The adoption of Act 82 affirmed this approach, and identified a range of new measures which are detailed below.

Figure 1. Pennsylvania Teacher Evaluation Formula: Beginning 2013-14 school year[2]



[2] PSSA- Pennsylvania Standardized State Assessment; PVAAS- Pennsylvania Value-Added Assessment System

Classroom observations, the basis of the former evaluation framework, will now account for 50 percent of a teacher's rating. Twenty percent is comprised of elective data: district officials are able to draw from a set of suggested measures that focus on additional measures of student achievement, such as student projects or standardized tests. The final thirty percent—15 percent teacher-specific data and 15 percent building-level data—relies primarily on student achievement as measured by the state's battery of standardized tests. Specifically, the building data will be derived from the new School Performance Profiles that were released earlier this month.

Much like the state's pilot program, the elements of the teacher evaluation system are being phased-in over a three-year period. This is due in part to the state using a three-year rolling average to calculate a teacher's impact on student learning.[xi] For example, for the 2013-14 school year, a teacher's rating will be derived from observation scores and building-level data, though districts have the option of factoring in elective data as well.

Pittsburgh Public Schools' Evaluation Formula

Aided by $90 million in public and private funding, Pittsburgh Public Schools (PPS) enacted its *Empowering Effective Teachers* initiative in 2007 through an agreement with the Pittsburgh Federation of Teachers (PFT), the local teachers' union. This work has focused on identifying alternative measures to recognize teacher effectiveness, piloting these measures across the district, using results to highlight teaching excellence, and executing improvement plans for professionals who may be struggling.

Under this initiative, PPS participates in the Measures of Effective Teaching (MET) study, funded by the Bill and Melinda Gates Foundation. The study sponsored partnerships between teachers and educational organizations in seven cities nationwide[3] with the goal of determining the most reliable ways to assess teacher effectiveness. The study confirmed earlier findings that multiple evaluative measures are more reliable than any single indicator. Specifically, when teachers in the study were rated using classroom observations combined with value-added scores—which are derived from statistical formulas that estimate an individual teachers' contributions to student learning—and student surveys, the results were more predictive of student achievement than classroom observations alone.

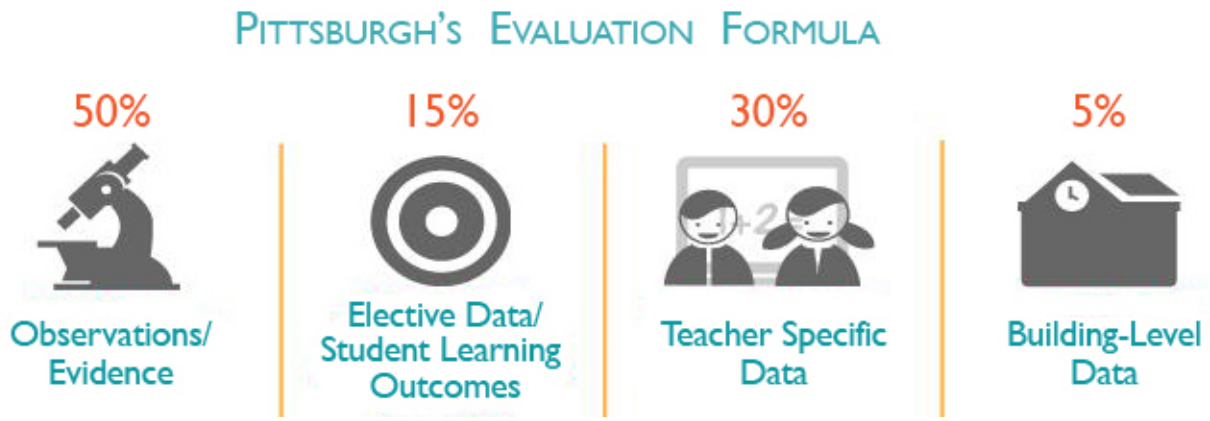
Like the newly-enacted state policy, the Pittsburgh model bases 50 percent of a teacher's rating on classroom observations. But under a one-year waiver from the Pennsylvania Department of Education, Pittsburgh's model adjusts weights for the other three categories:

1. **Building-level data** is reduced from 15 percent to **five percent**;
2. **Elective data** composes **15 percent** instead of 20 percent in the state model; and
3. **Thirty percent** of the PPS rating system will come from **teacher-specific data**.

Figure 2, below, shows the evaluation formula that will be used in Pittsburgh Public Schools during the 2013-14 school year.

[3] Charlotte-Mecklenburg Schools, NC; Dallas Independent School District, TX; Denver Public Schools, CO; Hillsborough County Public Schools, FL; Memphis City Schools, TN; New York City Department of Education, NY; and Pittsburgh. The MET project is a large-scale, multi-year study; this data was collected between 2009 and 2011.

Figure 2. Pittsburgh Public Schools' Teacher Evaluation Formula



Within these categories, Pittsburgh Public Schools has developed customized tools to support implementation of its approach to measuring teacher effectiveness:

- **Elective Data:** The district will base 15 percent of a teacher's ratings on student survey results that record evaluations of the teacher using the Tripod survey. This tool, utilized in the MET study, measures student responses to statements regarding classroom climate and learning environment—qualities that cannot be isolated through other indicators, such as value-added measures.

  Wilkerson, et al. (2000) found that students' ratings of teachers had a higher correlation with student achievement on reading and language arts tests than teachers' self-ratings and a higher correlation with student achievement on math, reading, and language arts than principal evaluations.[xii] The MET project reported that the top quartile of teachers based on student survey results saw their students gain an additional 4.6 months of math in a year (based on standardized test scores) when compared to teachers in the bottom 25 percent.[4][xiii]

  Additionally, student survey results from the MET study were more reliable than either classroom observations or value-added measures. When comparing several student surveys for a single teacher, the results were more consistent than multiple results of the other indicators.[xiv]

- **Teacher Specific Data:** Pittsburgh also created its own value-added model, electing not to use data from the Pennsylvania Value-Added Assessment System (PVAAS). The primary difference between the state model and Pittsburgh's in terms of teacher-level data is the explicit effort to control for student-level characteristics (*e.g.*, English Language Learner status, participation in gifted or AP courses, and various physical and learning disabilities). While value-added modeling has the capability of adjusting for student-level characteristics, PVAAS does not explicitly incorporate such traits[xv] However, the PVAAS model does include all prior assessment data on students, offering the potential for alternative controls.

[4] In performing this analysis, the study controlled for student demographic characteristics and prior academic performance.

Implications for state policy and local practice

Pennsylvania's new policy on teacher evaluation only begins the process of changing instructional practice. Implementing the policy effectively across 500 districts and aligning it with other reforms—including the state's new system of Common Core-infused academic standards—represent major challenges for education leaders at the state, district, and building levels.

Below are key points for consideration based on a review of the current research:

TEACHER OBSERVATION: Research has demonstrated that observation-based evaluation can have "a substantial positive relationship with student achievement and that the instructional practices measured by these systems contribute to student learning" when accompanied by focused professional development.[xvi] However, the efficacy of observations is predicated on a reviewer who can identify good teaching, and rigorous training in the use of the evaluation tools.[xvii] However, cuts to administrative capacity in many districts may complicate the implementation and execution of thorough, regular teacher observation.

The use of **STUDENT PERFORMANCE DATA**, will be primarily based on standardized test results. Using state assessment data in teacher evaluations is complicated by the fact that a majority of teachers work in untested subjects such as art, music, or social studies—areas described as "The Other 69 Percent" by the Center for Educator Compensation Reform.[xviii] Another layer of complexity involves the implementation of curricular and assessment changes associated with the Pennsylvania Core Standards and the Keystone Exams, which are being implemented alongside the new evaluation system.[xix]

In both the statewide and Pittsburgh formulas, a form of **VALUE-ADDED MODELING (VAM)** will be used to measure student growth. Proponents of VAM believe it is a useful objective tool for measuring and identifying teacher effectiveness, others cite reasons that decisions stemming from VAM may be flawed.

As previously stated, one difference between the statewide value-added model and Pittsburgh's is the inclusion of student-level characteristics in the data set. Researchers have debated whether value-added models need to take this step. Choi, Goldschmidt, and Yamashiro (2006) tested how controlling differences in socioeconomic status (SES) on both student- and school-level data impacts calculations.[xx] They reported that value-added measures do not necessarily need to account for differences in student characteristics, as they are already captured in the students' initial status. In other words, using all available state standardized test scores as a student's starting-out point already captures SES and other characteristics that correlate with academic performance.

Other researchers have questioned this assumption, especially when considering classroom composition and its relationship with student performance. In their review of assessment data from the California Standards Tests, Newton, Darling-Hammond, Haertel, and Thomas (2010) found that a teacher's ranking was significantly correlated with the demographic composition of their classroom; for instance, teachers with higher proportions of traditionally-disadvantaged students were negatively affected, while the opposite held true for high proportions of historically high-achieving students.[xxi] This finding was replicated in models that did and did

not account for student demographics. The authors conclude that this suggests either "teaching greater proportions of more advantaged students may have been advantaged in their effectiveness rankings, or that more effective teachers were generally teaching more advantaged students."

A second area of debate concerns the potential for low year-to-year correlations of teachers' value-added scores.[xxii] Two studies summarized in Sass (2008)—which examined assessment data from California and Florida—observed that between 10 and 15 percent of the highest-ranked (by quintile) teachers according to value-added scores fell to the lowest quintile the following school year; the reverse also held true.[xxiii]

It should be noted that different value-added/growth models yield different results in terms of reliability and relationship to student characteristics. The standard PVAAS approach for teachers yields repeatability estimates around 0.70 to 0.80 for three-year estimates. For this reason, Pennsylvania incorporates three years of value-added data in its teacher accountability system. As with multiple measures generally, additional data may help mask instability in a smaller data set but it is "hard to guarantee or even be reasonably sure" that a teacher's ranking may be based on "unexplainably low performance for two or three years in a row."[xxiv]

Conclusion

Although teacher evaluations in Pennsylvania's public schools were previously based solely on observations by administrators, several new components are being developed under Act 82. Whether in Pittsburgh or statewide, the policy and practical considerations inherent in teacher evaluations—and the difficulty of applying standardized measures for highly-individualized professional roles—represent new challenges. It is likely that the legislature, state education department officials, and local education leaders will need to continually revisit elements of this policy to ensure successful implementation. Drawing on lessons learned will be especially vital for the roll out of new evaluation systems for principals and nonteaching professionals in 2014-15.

Works Cited

[i] National Council on Teacher Quality. (2012). State of the states 2012: Teacher effectiveness policies. Retrieved from http://www.nctq.org/dmsView/State_of_the_States_2012_Teacher_Effectiveness_Policies_NCTQ_Report.

[ii] New Jersey Department of Education. (2013). Achieve NJ: Teacher Evaluation and Support in 2013-14. Retrieved from http://www.state.nj.us/education/AchieveNJ/intro/1PagerTeachers.pdf.

[iii] Research for Action. (2011). Teacher Effectiveness: The National Picture and Pennsylvania Context. Retrieved from http://www.researchforaction.org/wp-content/uploads/2011/09/RFA-PACER-Brief-Teacher-effectiveness-Sept.-2011.pdf.

[iv] McMurrer, J and Yoshioka, N. (2013). States' perspectives on waivers: relief from NCLB, concern about long-term solutions. *The Center on Education Policy*. Retrieved from: http://www.cepdc.org/displayDocument.cfm?DocumentID=418

[v] *Ibid.*

[vi] U.S. Department of Education. (2010). Race to the top scoring rubric. Retrieved from http://www2.ed.gov/programs/racetothetop/scoringrubric.pdf.

[vii] Lomax, E. and Kuenzi, J. (2012). Value-Added modeling for teacher effectiveness. *Congressional Research Service*. Retrieved from http://www.fas.org/sgp/crs/misc/R41051.pdf.

[viii] Mezzacappa, D. (2013, July 22). District to release report on cheating investigation at 19 schools. *The Philadelphia Public School Notebook*. Retrieved from http://thenotebook.org/blog/136221/district-release-report-cheating-investigation-19-schools.

[ix] The Danielson Group. (n.d.) The framework for teaching. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching.

[x] Heneman, H., Milanowski, A., Kimball, S., and Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay. *Consortium for Policy Research in Education*: pp. 4-5 Retrieved from http://eric.ed.gov/PDFS/ED493116.pdf.

[xi] Pennsylvania Department of Education (August, 2013) PVAAS teacher specific reporting guide to implementation. SY2013-14 Statewide Implementation: Pennsylvania Value-Added Assessment System.

[xii] Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal and self-ratings in 360° feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education*, 14(2), 179–192. Retrieved from: http://cfl.ctu.edu.vn/learningresource/ebooks/49.pdf

[xiii] MET Project. (2012). Asking students about teaching. Retrieved from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.

[xiv] *Ibid.*

[xv] Pennsylvania Department of Education. (2013). Frequently asked questions (FAQs): Pennsylvania Value-Added Assessment System (PVAAS). Retrieved from https://pvaas.sas.com/unrestricted.download?ab=dn&as=a&yq=2&wy=PVAASPilot_FAQs%20Teacher%20Specific%20Reporting.pdf.

[xvi] Heneman, H., Milanowski, A., Kimball, S., and Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay. *Consortium for Policy Research in Education*: pp. 4-5 Retrieved from http://eric.ed.gov/PDFS/ED493116.pdf.

[xvii] Goe, L., Bell, C., and Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. *National Comprehensive Center for Teacher Quality*: p. 8. Retrieved from http://www.tqsource.org/publications/EvaluatingTeachEffectiveness.pdf.

[xviii] Price, C., Schuermann, P., Guthrie, J., Witham, P., Milanowski, A., and Thorn, C. (2009). The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades. Retrieved from http://www.cecr.ed.gov/guides/other69Percent.pdf.

[xix] Hernandez, J. (2013, August 4). Results of new testing standard could complicate bloomberg's final months. ***The*** *New York Times*. Retrieved from http://www.nytimes.com/2013/08/05/nyregion/results-of-new-testing-standard-could-complicate-bloombergs-final-months.html?partner=rss&emc=rss&_r=1&.

[xx] Choi, K., Goldschmidt, P., & Yamashiro, K. (2006). Exploring models of school performance: From theory to practice (CSE Rep: No. 673). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

[xxii] Baker, B., Oluwole, J., and Green, P. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*. Vol. 21, No. 5. Retrieved from http://epaa.asu.edu/ojs/article/view/1298/1043.

[xxiii] Sass, T. R. (2008). The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy. Policy Brief 4. Washington, D.C.: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research. Retrieved from: http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf

[xxiv] Baker et al. (2013)