Tulsa Public Schools' Teacher Observation and Evaluation System: Its Research Base and Validation Studies

Summary

The Tulsa teacher evaluation model was developed with teachers, for teachers. It is based on current, best practices and national research findings. Tulsa Public Schools has subjected its model to independent validation studies in both a no-stakes and higher-stakes context using working principals with only minimal calibration training. The studies confirmed that the Tulsa teacher evaluation model measures teacher practices that track student achievement growth. By responding appropriately to the research findings and input of working teachers and principals, Tulsa Public Schools is ensuring that it has an empirically robust system that teachers, administrators, parents and other stakeholders trust.

Research-Based and Teacher-Developed

Developed with teachers through intensive study of research and best practices

Tulsa Public Schools began the development of its evaluation system in 2009 as part of its education reform work with the Bill and Melinda Gates Foundation. A study group comprised of national evaluation experts, TPS teachers, curriculum specialists and principals reviewed dozens of teacher evaluation instruments and research studies. Using the research findings and their professional expertise, the study group developed recommendations and a list of specific principles to guide the overall structure and substance of the teacher evaluation rubric. A smaller team created from the members of the work group used the guidance and the underlying research materials to create the evaluation framework (the evaluation rubric).

Research base

The research base supporting the TPS framework is broad in that it includes the work of multiple practitioners and academic researchers. Two groups of studies, however, are particularly noteworthy: the recommendations of the Northwest Regional Educational Lab[1] and the research findings of Harvard researcher Thomas Kane and his colleagues.[2] These studies confirm that the underpinnings of the Tulsa model are observable practices associated with increases in student achievement.

[1] Kathleen Cotton, Northwest Regional Educational Lab (2000). "The Schooling Practices that Matter Most." ASCD.

[2] Kane, Thomas J., Taylor, Eric S., Tyler, John H., and Wooten, Amy L. (2011). "Identifying Effective Classroom Practices using Student Achievement Data," The Journal of Human Resources, 46:3. See also

In 2000, ASCD (the Association for Supervision and Curriculum Development) published a well-regarded paper by Kathleen Cotton and the Northwest Regional Educational Lab that provides valuable insight into what should be included within a teacher evaluation framework. Their publication analyzed research findings on educational practices to identify the core contextual and instructional factors that enable students to learn successfully. Not surprisingly, many of the attributes noted in Cotton's paper relating to teacher practices and competencies were well-established characteristics of effective teaching and continue to be so. Indeed, in addition to the Tulsa model, the practices are commonly found within many well-known teacher frameworks and education treatises, including, but not limited to, Charlotte Danielson's *Framework for Teaching* and Robert Marzano's *The Art and the Science of Teaching*.

With regard to Tulsa's model, specifically, its rubric assesses many of the contextual factors identified in the Cotton paper, including the teacher's ability to clearly communicate and support high behavioral expectations, to consistently apply rules and standards of behavior, to stop disruptions quickly, maximize learning time, differentiate and adapt instruction to the needs of faster and slower learners, pace lessons appropriately, minimize time for transitions, monitor student progress, etc.

The Tulsa model also incorporates many of the instructional practices identified as vital to increasing student achievement. Among other factors, Tulsa's rubric measures a teacher's ability to explain lessons and objectives clearly, to describe the relationship of the current lesson to previous learning, to use strategies such as advance organizers, to ask questions that engage student interaction and enable the teacher to monitor student understanding, to provide for "wait time" when questioning students, and give timely feedback and reinforcement.

Many of the practices incorporated within the Tulsa model are also proven in empirical terms by published, peer-reviewed research. A research team led by Thomas Kane, an economist with Harvard Graduate School of Education, analyzed numerous teacher practices and whether a teacher's proficiency in using a specific practice tracked his or her quantitative impact on student achievement growth (i.e., whether the teacher's observation score on certain performance criteria tracked that teacher's value-added score). The researchers found that a teacher's competence in certain practices did, in fact, predict the achievement gains made by the teacher's students in both math and reading. These practices, derived primarily from the descriptions in Charlotte Danielson's *Framework for Teaching*, centered on matters of classroom management and instructional effectiveness. For example, the practices included, among others, the teacher's ability to manage and monitor student behavior and respond appropriately, as well as the teacher's ability to use higher-order questioning techniques and provide timely feedback to student about their progress.

Kane, Taylor, Tyler, and Wooten. (2010). "Identifying Effective Classroom Practices Using Student Achievement Data," National Bureau of Economic Research Working Paper 15083. Kane, Taylor, Tyler, and Wooten. "Evaluating Teacher Effectiveness," Education Next. www.educationnext.org/ evaluating-teacher-effectiveness. Summer 2010.

Tulsa's evaluation framework incorporates the practices Kane found to be associated with student achievement. In particular, Tulsa's model measures a teacher's ability to: clearly define and support expected behavior; develop plans to achieve identified objectives; use higher-level questioning techniques; engage all learners; differentiate instruction and activities to respond to differences in student needs; provide adequate and timely feedback; adjust instruction based on the results of monitoring; and create a caring, respectful and effective learning environment.

Validation Studies

A validation study determines if the evaluation protocol measures what matters—whether teachers' individual evaluation scores as measured by a qualitative evaluation instrument track their quantitatively measured impact on student learning. As the American Institutes for Research explains, a validation study of an evaluation protocol/instrument should measure the "correlation between a teacher's evaluation protocol score and the teacher's value-added score."[3] Tulsa has subjected its evaluation system to two types of validation studies—a rigorous study conducted through the Bill and Melinda Gates' MET Validation Engine project as well as a correlational analysis of Tulsa's own, "real-world" evaluation and value-added data by the University of Wisconsin's Value-Added Research Center. Both independent studies validated the Tulsa model.

MET Validation Engine Analysis

In the fall of 2011, Tulsa Public Schools participated in the pilot of the MET Validation Engine—a research project of the Bill and Melinda Gates Foundation developed by Empirical Education Inc., an education research company. The Validation Engine Project allowed the District to determine the predictive validity and rater consistency of the Tulsa model's protocol—its teacher evaluation rubric—through an independent study conducted by national experts.

Using a web-delivered software tool, a representative sample of Tulsa principals viewed over 160 classroom observation videos and rated those videos using the Tulsa teacher evaluation rubric. The videotaped lessons were full recordings of actual ("real-world") math and English/Language Arts classes from other school districts around the country and ranged in length from 45 minutes to 1.5 hours. The researchers from Empirical Education had several years of value-added data for each teacher whose classroom performance was viewed and ranked by Tulsa's principals, but this information was not shared with the Tulsa principals, who had to judge the teacher's performance based solely on their use of the Tulsa model's evaluation rubric. By comparing the principals' rankings with the value-added

[3] It is inappropriate for validation purposes to compare teachers' evaluation scores with student or school attainment scores—measures of proficiency/achievement calculated outside the context of complex growth modeling. To do so ignores the fact that students have drastically different levels of prior achievement (starting points) at the beginning of a school year and that student achievement is also affected by individual student characteristics unrelated to a teacher's practices and competencies.

scores of the teachers, the researchers from Empirical Education were able to test the validity of the Tulsa model. Specifically, they worked to determine whether, and to what extent, the observation instrument captures and reflects teacher practices that are correlated with growth in student achievement.

A notable component of this study is that it used working principals with very minimal calibration training—not expert raters of small research teams. As explained in the recent research paper by the Gates Foundation titled *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*,[4] when the study of an observation instrument uses research teams of the instrument developers, themselves, "it can be hard to distinguish between the power of the instrument and the special expertise of the instrument developers themselves to discern effective teaching."[5] Stated another way, the instrument needs to be transferable. "We don't just want to know whether a small group of experts can distinguish between effective and ineffective instruction; we want to know whether a larger group of observers with little special expertise beyond a background in teaching can be trained to look for the same competencies."[6]

The findings of the MET Validation Engine study were positive and confirmed that the Tulsa model measures what matters—that it captures practices that are empirically associated with gains in student achievement. Specifically, the study revealed that every indicator included within the Tulsa model that a principal uses when observing a classroom performance is positively correlated with growth in student achievement as measured by state assessments.

Analysis by the University of Wisconsin

In addition to the MET Validation Engine Project, the Tulsa model has also been studied by the University of Wisconsin's Value-Added Research Center (VARC). Instead of evaluating the Tulsa rubric in the context of isolated classroom observations, this research team studied Tulsa's evaluation system by comparing teachers' value-added data to their respective overall evaluation scores—which are based largely on classroom observations *but also* the totality of the principals' experience with the teacher throughout the evaluation period, including competencies that are not observable in a classroom observation such as leadership qualities and attention to professional growth and development. This study used actual evaluation and value-added data from the District, itself. As such, this analysis allowed researchers to study the use of the evaluation system in a real-world, high-stakes setting—an important test of validity.

To conduct the study, the researchers from the Value-Added Research Center needed teachers' value-added scores and those teachers' respective overall evaluation scores. Tulsa Public Schools has value-

[4] *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains*. Bill and Melinda Gates Foundation. Lead Authors: Kane, Thomas J.; Staiger, Douglas O., 2012. http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf.
[5] *Id.* at p.5.
[6] *Id.*

added scores for the 2010-2011 school year for all teachers in subjects for grades 4-12 for which there were state assessments. Because the Tulsa teacher evaluation model has been used District-wide since 2010-2011, it also has a database of teachers' overall evaluation scores as measured by the Tulsa model for that same time period. The VARC research team calculated the correlations between a teacher's evaluation score using the Tulsa model and his or her value-added score for the 729 instances in which there were both types of data. The researchers also determined which indicators were more predictive of student achievement growth than others.

As with the MET Validation Engine, the research team from VARC issued positive results validating the Tulsa model. Teachers' overall evaluation scores as measured with the Tulsa evaluation model were positively correlated with their respective value-added scores. Similarly, every indicator in the Tulsa model was positively correlated with this student growth measure. Indeed, the average correlation between the teachers' value-added scores and their respective evaluation scores across all subjects using the Tulsa evaluation system was 0.22. The largest samples were those for fourth and fifth grades. The correlation for fourth grade math was 0.23 and the correlation for fifth grade math was 0.45. The equivalent numbers for reading were 0.20 and 0.18.

Overall, these results are similar to those described in academic literature of well-known evaluation instruments.[7] For example, in the 2010 study noted above by Kane et. al., "Identifying Effective Classroom Practices using Student Achievement Data," regarding a nationally recognized evaluation model, the researchers found an overall correlation between value-added scores and a observation-based scores for math of 0.17 and an overall correlation for reading of 0.21. The Kane study also found the items measuring classroom management and instruction are most highly correlated with value-added. Correlations of Tulsa data have the same result. Notably, the results also mirrored to a significant extent the findings of the MET Validation Engine pilot with regard to which indicators were good predictors of value-added scores.

[7] At first, one might expect correlations above 0.20, but the academic literature consistently finds estimates in this range for three important reasons. First, a teacher's value-added score is a statistical estimate of their true value-added score. Plus, the observation score is an estimate of the true observation score of what a master grader would find if observing every class for the entire year. Finally, we do not expect the true value-added score to be perfectly correlated with the true observation score because they are different measures of effectiveness. When all three of these factors are combined, it drives down the correlation between the value-added score and the qualitative evaluation score one would expect to a correlation that is below 0.5 yet still positive. This is what one sees empirically in both Tulsa and the academic literature.

For related discussions and similar findings in a slightly different context, see *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains,* supra, in which researchers analyzed teacher performance and student growth data relevant to nearly 1500 teachers to determine the alignment of several national teacher observation instruments and future value-added scores.

Using the Validation Data for Continuous Improvement

Both the MET Validation Engine and the University of Wisconsin/VARC studies provided rich details about the Tulsa evaluation protocol. The District will use this data in a variety of ways to enhance its evaluation system. For example, in the MET Validation Engine study, the indicators in the Tulsa model with the highest predictive power were those relating to a teacher's competence in monitoring her students' learning and modifying her instruction accordingly; planning lessons relative to short-term and long-term objectives based upon the results of monitoring; demonstrating and modeling the desired skill or process for her students; and summarizing the lesson. The findings issued by VARC confirmed the importance of these indicators and others.

The District will leverage the power of the more powerful indicators by intensifying the principal calibration training on them and ensuring that the rubric language relating to the indicators is as clear and precise as possible. Likewise, the District will reevaluate the language pertaining to less powerful indicators. For example, the indicator relating to a teacher's ability to optimize the classroom's physical learning environment was not a strong predictor in the MET Validation Engine pilot. While it was positively correlated with student achievement gains, it was only minimally predictive, especially in comparison to the predictive abilities of other indicators within the Tulsa framework. The same is true of the indicator relating to leadership, such as a teacher's willingness to contribute to school and district initiatives, a characteristic not observable in a classroom observation alone. The VARC research indicated that it is much less powerful than other indicators, and as such, the District will analyze its language and consider alternative language that would more closely track student achievement gains.

Conclusion

As noted above, the Tulsa evaluation model is unique in that it was developed with teachers, for teachers. It is also empirically sound. It is based on current, best practices and national research findings. Independent studies have validated and confirmed that the Tulsa model measures what matters. By appropriately responding to research findings and leveraging the strengths of its teacher evaluation rubric, Tulsa Public Schools is ensuring that it supports the best use of the teacher evaluation system—the identification and development of teacher practices that have the greatest impact on student achievement.