STRATEGIC DATA PROJECT

SDP FELLOWSHIP CAPSTONE REPORT

Analyzing and Improving Multiple Measure Teacher Evaluation Systems

Yueming Jia, Fort Wayne Community Schools
Todd Cummings, Fort Wayne Community Schools
Cara Jackson, Urban Teacher Center
Megan Clifford, Oklahoma State Department of Education
Stephen Hoch, Tulsa Public Schools

SDP Cohort 5 Fellows

Strategic Data Project (SDP) Fellowship Capstone Reports

SDP Fellows compose capstone reports to reflect the work that they led in their education agencies during the two-year program. The reports demonstrate both the impact fellows make and the role of SDP in supporting their growth as data strategists. Additionally, they provide recommendations to their host agency and will serve as guides to other agencies, future fellows, and researchers seeking to do similar work. *The views or opinions expressed in this report are those of the authors and do not necessarily reflect the views or position of the Center for Education Policy Research at Harvard University.*

Framing the Problem

Recent state and federal policies have focused a great deal of attention on teacher evaluation systems to improve instruction. For example, in 2012, President Obama granted waivers from No Child Left Behind (NCLB) requirements to several states based on their progress toward implementing reform measures that included the development and implementation of rigorous teacher evaluation systems. While teacher evaluation systems are not new in most states and districts, the implementation of several new multiple measure systems—as well as the increased stakes attached to them—have led to the need for more research regarding the reliability and validity of these measures and effective interventions to improve the models.

While well-designed and properly implemented teacher evaluation systems have the potential to provide administrators with accurate information about the true effectiveness of teachers over time, poorly designed or poorly implemented systems run the risk of providing inaccurate or inconsistent information with several detrimental effects, including wrongful termination or misdirected performance incentives. The need for timely information so that mid-year adjustments and support can be provided also makes the validity and reliability of teacher observation data particularly important as student growth data is not typically available until the end of the school year. Unfortunately, there is little practical guidance on what analyses can be conducted to assess the strengths and limitations of teacher evaluation systems and what steps can be taken to improve the accuracy of the evaluations.

This research presents findings from four education agencies that have implemented multiple measure teacher evaluation systems: Fort Wayne Community Schools (FWCS), Urban Teacher Center (UTC), the Oklahoma Department of Education (OKSDE), and Tulsa Public Schools (TPS). Each agency provides information on how to assess and improve specific aspects of the teacher evaluation system. FWCS's case study explores the district's initiatives to improve the inter-rater reliability of classroom observation ratings, including a train-the-trainer reliability training program and the development of a user manual of a classroom teacher effectiveness rubric. UTC's case study discusses an intervention to improve inter-rater reliability by having instructional coaches conduct paired observations. Additionally, it presents an approach to assessing how reliability is impacted by the number of observations and items on an observation rubric. OKSDE's case study explores the specifications of its value-added model (VAM) and its impact on fairness, understandability, precision and comparability. It also describes the relationship of its VAM results to observable school-level and teacher characteristics. Finally, TPS's case study reviews the process and policy tradeoffs related to designing and implementing a multiple-measures evaluation system in a large school district. Taken together, the purpose of these case studies is to provide state and local leaders with more information on the specific threats regarding validity and reliability that other agencies have identified and how agencies addressed these threats so that they can improve their own systems.

Literature Review

As illustrated in Figure 1, the theory behind teacher evaluation reform is that a teacher evaluation system enables agencies to identify low-performing teachers and to provide feedback and support to develop the teachers they retain, thus enhancing teacher effectiveness and generating student achievement growth. To succeed in this ambitious goal, an evaluation system must provide accurate information differentiating among teachers as well as information about individual teachers' specific strengths and weaknesses so that feedback and support can target areas of need. If the teacher evaluations are not based on valid and reliable measures, agencies risk miscategorizing teachers, providing invalid feedback, or inefficiently targeting support. In other words, the potential for teacher evaluation systems to improve educational outcomes for students depends on the quality and use of the underlying teacher evaluation data.

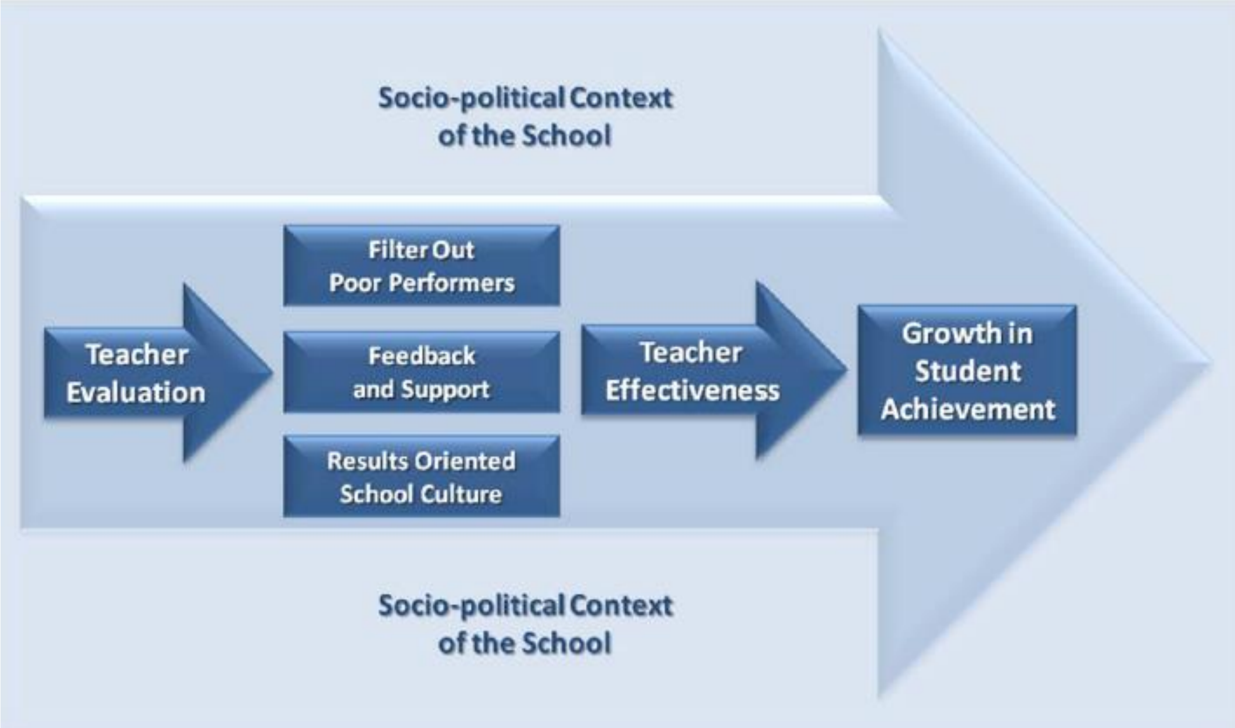

Socio-political Context of the School
Teacher Evaluation
Filter Out Poor Performers
Feedback and Support
Results Oriented School Culture
Teacher Effectiveness
Growth in Student Achievement
Socio-political Context of the School

Figure 1. Theory of Action Underlying Teacher Evaluation and School Improvement (Hallinger, Heck, & Murphy, 2014).

Recognizing the important impact teachers have on student outcomes but faced with the challenge that historical systems of teacher evaluation largely failed to distinguish between effective and ineffective teachers (Weisburg et al., 2009), policymakers have shifted more recently to the use of multiple-measures to improve the validity and reliability of teacher evaluation systems and to make better use of teacher effectiveness data. Evaluation systems based on multiple measures can be more reliable than evaluation systems based on a single measure; while more reliable than single measure systems, multiple measures evaluation systems are imperfect and will lead to classification errors where some teachers rated as ineffective are in fact effective teachers and vice versa (Goldhaber& Loeb, 2013).

Observations

Of the measures used in teacher evaluation, classroom observations have an advantage over other potential evaluation metrics in that all teachers can be observed; in addition, classroom observations can be used for both formative and summative purposes. Observations have the potential to serve a formative purpose by supporting mid-year corrections in teaching practice, to the extent that information from classroom observations is immediately available and readily understood by teachers. In Taylor & Tyler's (2012) study of mid-career math teachers in the Cincinnati Public schools, the authors find evidence that quality classroom-observation-based evaluation and performance measures can improve mid-career teacher performance both during the period of evaluation and in subsequent years. The finding that evaluation improves performance in subsequent years is consistent with the theory that evaluation can serve as an investment in human capital. Classroom observations can even support mid-year corrections in teaching practice. Summative data from classroom observations can also provide information to guide personnel decisions.

While classroom observations have considerable potential to foster improvements in teaching, observation rubrics can be used effectively or ineffectively; reliability and validity are functions of the users of the tool, as well as of the tool itself (Sartain, Stoelinga, & Brown, 2011). As agencies implement new and refined observation protocols, they will need to assess inter-rater reliability and whether the tool used meaningfully differentiates across the spectrum of teacher effectiveness. MET project researchers have explored a variety of ways to enhance the reliability of classroom observation data specifically through the use of video and calibration protocols. Ho and Kane (2013) examined classroom observation data from 129 raters (53 school administrators and 76 peer raters) who observed lessons from 67 teachers to explore the implications of different approaches to obtaining high levels of accuracy and reliability in classroom observations. They note that one can increase reliability without increasing the number of total observations by using more than one observer for a given observation.

Value-Added Models (VAMs)[1]

VAMs are intended to capture teacher effectiveness by comparing how much a particular teacher improves student achievement (based on standardized assessments) to how much the typical teacher would have improved student achievement. In general, research indicates that there is important variation in teacher effectiveness that has educationally

[1] This report does not discuss Student Growth Percentiles (SPGs) because none of the case study agencies use SPGs in their teacher evaluation systems.

significant consequences for student achievement, and that VAM measures are likely to contain real information about teacher effectiveness that could be used to inform personnel decisions and policies (Corcoran & Goldhaber, 2013). The advantage of such measures is that in contrast to teacher observations, which often fail to distinguish among high and low performers, VAMs by design differentiate among teachers. However, many researchers have urged caution when using VAMs in high-stakes evaluation systems due to potential bias and imprecision in these estimates (e.g., ASA, 2014). Despite general agreement on the statistical properties of VAMs, the use of these measures in high-stakes teacher evaluation systems is quite contentious, in part because little is known regarding how educators might respond to high-stakes uses of such measures.

Student Surveys

Asking students about teachers to gain insight into teacher performance and to provide feedback–a common practice in universities across the nation—is now being used by states and local education entities in grades K–12. Student surveys provide actionable information for teachers (as do observations) and are related to student growth (as measured by value-added; see Kane &Staiger, 2012). Of the three measures of teacher effectiveness discussed, student surveys have the smallest body of research; however, the available research indicates that the new tool for measuring teacher effectiveness shows much promise.

In Balch's (2012) study on student surveys and their relation to teacher practice, student surveys were found to be more predictive of student achievement than teacher self-ratings, principal ratings and principal summative evaluations. Furthermore, Kane and Staiger (2012) found that student surveys were not only more predictive of student growth than observations but were more reliable than both observations and value-added. Both research papers cited above have used results from the Tripod Student Survey, the most widely known student survey for K–12 classrooms. Given the success and rapid adoption of student surveys as a tool for teacher feedback and evaluation around the nation, a number of new entrants have entered the market. The research and promising evidence related to the Tripod Student Survey has not yet been tested with the newer survey instruments.

Using Multiple Measures:

Whether the theory of action behind teacher evaluation systems is viable depends on whether these systems accurately differentiate among teachers and provide useful feedback in a timely manner to facilitate improvements in teaching practice. Observation-based evaluation data is highly actionable and can differentiate teacher practice when implemented in an effective manner. Historically, however, observation based rating systems have failed to identify low performing teachers; as many as 99% of teachers received evaluations ratings of satisfactory or better (Weisburg et al., 2009). Value-added data does not provide teachers with information regarding the strengths and weaknesses in their teaching practice, but it can be used to differentiate among teachers. Student survey data provides teachers with formative feedback on how their students perceive them; however, relatively little research has been done on surveys as a measure of teacher effectiveness. Ultimately, combining multiple measures should provide a more accurate view of teacher effectiveness, but the degree of accuracy depends on the reliability and validity of the underlying components.

After selecting the measures that will be used in a multiple measures evaluation system, State Education Agencies (SEAs) and Local Education Agencies (LEAs) must decide how the measures will be combined to create a composite or summative rating of teacher effectiveness. The empirical evidence related to how much weight should be placed on each measure is minimal. In the MET Project policy brief, *Ensuring Fair and Reliable Measures of Effective Teaching,* Cantrell and Kane discuss the tradeoffs between reliability and predictive power when selecting weights in a multiple measures evaluation system. In their brief, Cantrell and Kane identify evidence that suggests value-added should account for 33%– 50% of the weight in a composite score with principal observation ratings making up no less than half of the remaining weight (Cantrell & Kane, 2013).

Case Study: Fort Wayne Community Schools

Agency Profile

Fort Wayne Community Schools (FWCS) is Indiana’s largest school district with nearly 2,000 teachers and over 32,000 students from pre-k to 12th grade. The district is diverse: the student body consists of 45% White, 24% Black, 16% Hispanic, 9% Multi-racial, 5% Asian, and 1% American Indian students. Additionally, 71% students are eligible to receive free or reduced-price meals.

In the 2012–13 school year, FWCS implemented a new teacher evaluation system. Under this system, classroom observation ratings determine 60% of teacher effectiveness scores. To evaluate teachers' instructional practices, FWCS uses the Classroom Teacher Effectiveness rubric, a state-developed model based on Charlotte Danielson's Framework for Teachers. The rubric consists of 24 measures covering four major domains: Purposeful Planning (5 items), Effective Instruction (9 items), Teacher Leadership (5 items) and Core Professionalism (5 items). The measures of the first three domains are rated in a 4-point Likert scale, ranging from ineffective (1) to highly effective (4). The Core Professionalism measures are binary: doesn't meet standard (0) and meets standards (1).

Policy/Research Questions

Prior to its full implementation, the district piloted the Classroom Teacher Effectiveness rubric in 2011–12. In the district, principals mainly conduct classroom observations in each school. While principals received training during the pilot year, they did not receive any additional training in subsequent years nor were new principals trained. This led to a gap in training when the inter-rater reliability project was launched. In addition, the FWCS teacher effectiveness rubric that guides principals' observations contained many vaguely defined terms. For these reasons, inter-rater reliability is a major concern of the teacher evaluation program.

To understand how well the district has evaluated teachers' instructional practices using the rubric and how FWCS can improve the reliability of classroom observations, the two SDP Fellows conducted two research projects on the district's classroom observation rating system to examine the following: 1) Validity of the district's classroom rating; and 2) Development of the reliability training program.

Project Scope and Timeline

The SDP Fellows began the program with a good understanding that their work in the district would focus on the district's Human Capital Management initiatives and in particular, the teachers' performance-based compensation system. After initial examination of the district's current teacher evaluation policy and interviews with district's leadership, the fellows identified that validity and reliability of classroom observation rating in the district was in the most urgent need of research and action. Accordingly, two projects were set-up addressing classroom observation ratings. Project 1 was a set of analyses on the validity of the current classroom observation rating. Project 2 targeted the development of a training program on inter-rater reliability.

Project 1: Analyses of Validity of FWCS's Classroom Observation Rating: This analysis utilized teacher and student data from the 2013–14 school year as this was the most complete year of data. The analyses focused on classroom observation ratings, which were also linked to student growth data.

The analyses sought to understand how well the district's classroom observation rating differentiated teachers' performance within the district, within buildings, between experienced teachers and first-year teachers, and between teachers with a Master's degree and teachers with a Bachelor's degree. FWCS also investigated the relationship between classroom ratings and students' academic growth. For a detailed outline of the steps taken for this project, please see the FWCS Appendix.

Project 2: Development of FWCS Inter-Rater Reliability Training Program: In Project 2, the fellows developed and documented the district's proposed initiatives to improve the inter-rater reliability of classroom observation rating via a train-the-trainer reliability training program and develop a user manual for the classroom teacher effectiveness rubric. There were four deliverables for the project:

1. A plan for addressing inter-rater reliability and measurement;
2. A process for addressing inter-rater reliability and measurement;
3. A sustainable statistical model for measuring rater agreement;
4. A sustainable, on-going professional learning/support model.

To produce these deliverables, the fellows, LEA leaders and a third-party consultant completed calibration using a certification calibration engine and produced locally developed videos showing highly effective teachers. For a detailed outline of the steps taken for this project, please see the FWCS Appendix.

Results/Impact

The Validity of FWCS's Classroom Observation Rating: Between 2012–13 and 2013–14 school years, over 55% of FWCS teachers were rated effective, around 40% were rated highly effective, and about 3% were rated as improvement necessary or ineffective based on FWCS's classroom observation rubric. The ratings identified differences in teacher effectiveness between first-year teachers and experienced teachers. The result of multilevel logistic regression analysis showed that first-year teachers were significantly less likely to be in the highly effective category than experienced teachers in the district (Z_{2013} =-5.10, p<.001; Z_{2014} =-4.16, p<.001). The estimated likelihood of being in the highly effective category was about nine (2013) and six (2014) times higher for experienced teachers in comparison to first-year teachers.

The ratings also identified differences in teacher effectiveness between teachers with a master's degree and a bachelor's degree. The result of multilevel logistic regression analysis showed that teachers with a master's degree were significantly more likely to be in the highly effective category than teachers with a bachelor's degree (Z_{2013} =6.38, p<.001; Z_{2014} =4.30, p<.001). Teachers with a master's degree were 2 times (2013–14) more likely in the highly effective category than teachers with a bachelor's degree.

Additionally, the ratings predicted student growth scores. There was a statistically significant and positive relationship between teachers' classroom observation rating and their students' growth scores (B=.10, Z=2.72, p=.01). However, the magnitude of this relationship was small. A one point increase in the average teacher's classroom rating score was associated with an increase of .10 points in students' growth, on average (Figure 2).

Schools in the district varied significantly in the strength of the relationship between classroom observation ratings and student growth. Within-school variation on observation ratings also differed by schools in the district. The correlation between classroom observation ratings and student growth was as high as .30 in some schools, but it was a negative relationship for several other schools (Figure 3). In terms of within-school variation, it ranged from .20 points at some schools to close to .50 points at others.

Development of FWCS Inter-Rater Reliability Training Program: The purpose of this program was to establish a training and certification system for classroom teacher effectiveness raters. Five steps were planned (Figure 4): 1) develop a detailed and user-friendly manual of FWCS Classroom Teacher Effectiveness Rubric, 2) create of a training module for a group of trainers, 3) implement district-wide training and supporting module, 4)implement certification procedures, and 5) provide post-certification supporting and monitoring/checking.

Manual development: In October 2014, the cabinet invited a total of 11 people from the district form a team to develop a manual of FWCS classroom teacher effectiveness rubric. This team included seven principals, two district coaches, and two administrators in the curriculum department. Six full-day meetings were held from October 2014 to April 2015. The manual development started with identification of terms and words that might require further explanation and then a format for defining terms was developed. After that, a small group discussion was conducted to define the terms and provide examples. Once all the selected terms were defined and examples were provided, the whole group reviewed them together. At last, a working manual was distributed to all principals in the district for feedback. Revisions were made based on their suggestions. A draft manual was produced in the beginning of April 2015, which included a statement of purpose and theoretical framework for the rubric, clear definitions and examples of the rubric descriptors, mechanisms of the rubric, and general rules and procedures for use of eWalk with the rubric.

Train-the-Trainer: The manual development process was accompanied by training of the trainers. The participants of manual development became the first group of rater trainers in the district. In their meetings, using Empirical Education's Calibration and Certification Engine powered by the MET Study videos, all trainers watched a series of video clips together and rated the clips independently using the indicators in the rubric. Discussion about the ratings was conducted after each clip watching to reach an agreement until the team reached 80% agreement—the percent expected of all principals to become certified.

Perceived Initial Impact: According to the participants, the work of collegial conversations around definitions, construction of common meaning regarding instruction and the RISE rubric, and vertically aligned teams has already impacted their system of support for teachers. Participants have reported: 1) Increased precision and quality of feedback comments, 2) more consistent ratings across all forms of feedback, 3) Greater clarity and understanding in the relationship between Domains 1 and 2, and 4) Better understanding on the part of teachers and coaches of the terms and vocabulary in the RISE rubric.

Lessons Learned and Next Steps

The leadership team is adamant in sharing lessons with their colleagues. The next steps include building in longer time than expected to engage the administrators who are not on the leadership team in the same level of conversation and dialogue that the leadership team experienced. There was a great deal of concern that the process would be rushed and administrators would not have the same rich experience as the leadership team had. They appealed to Cabinet for longer working time and multiple windows to certify.

Figure 2. Relation of Classroom Ratings and Student Growth Scores

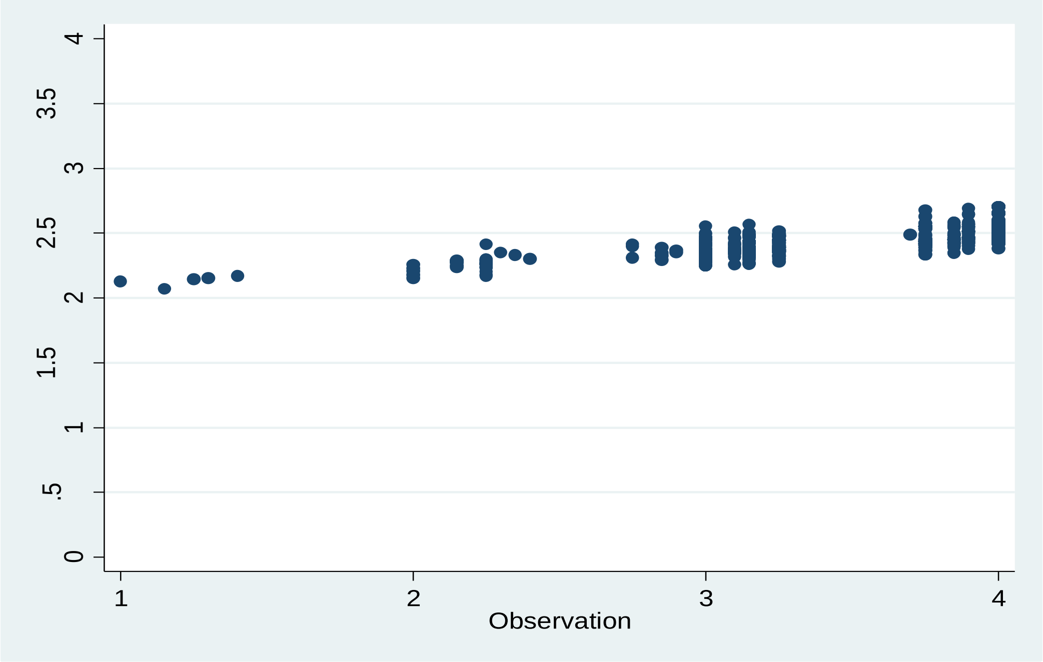


Figure 3. Relation of Classroom Ratings and Student Growth Scores by Schools

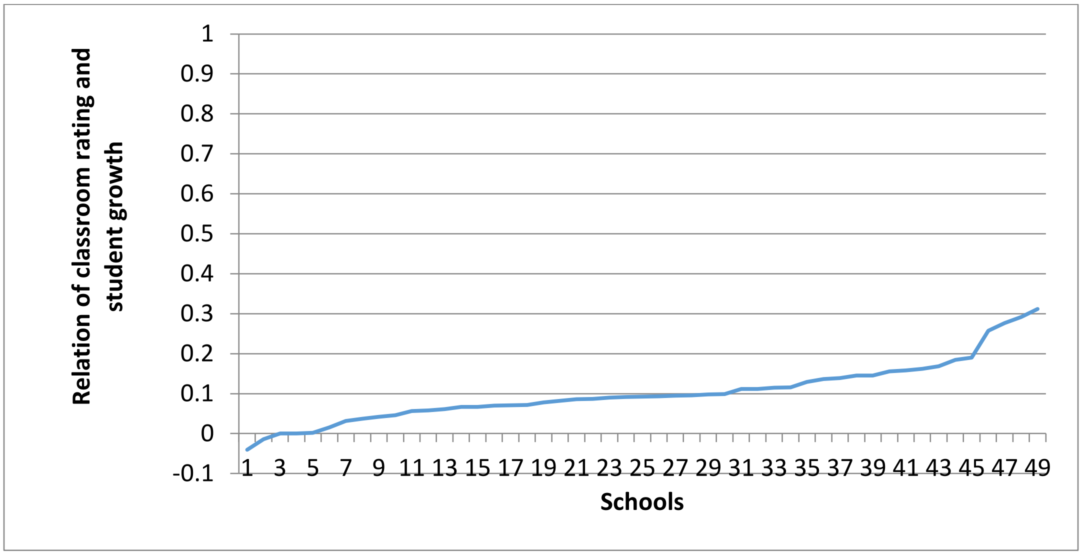
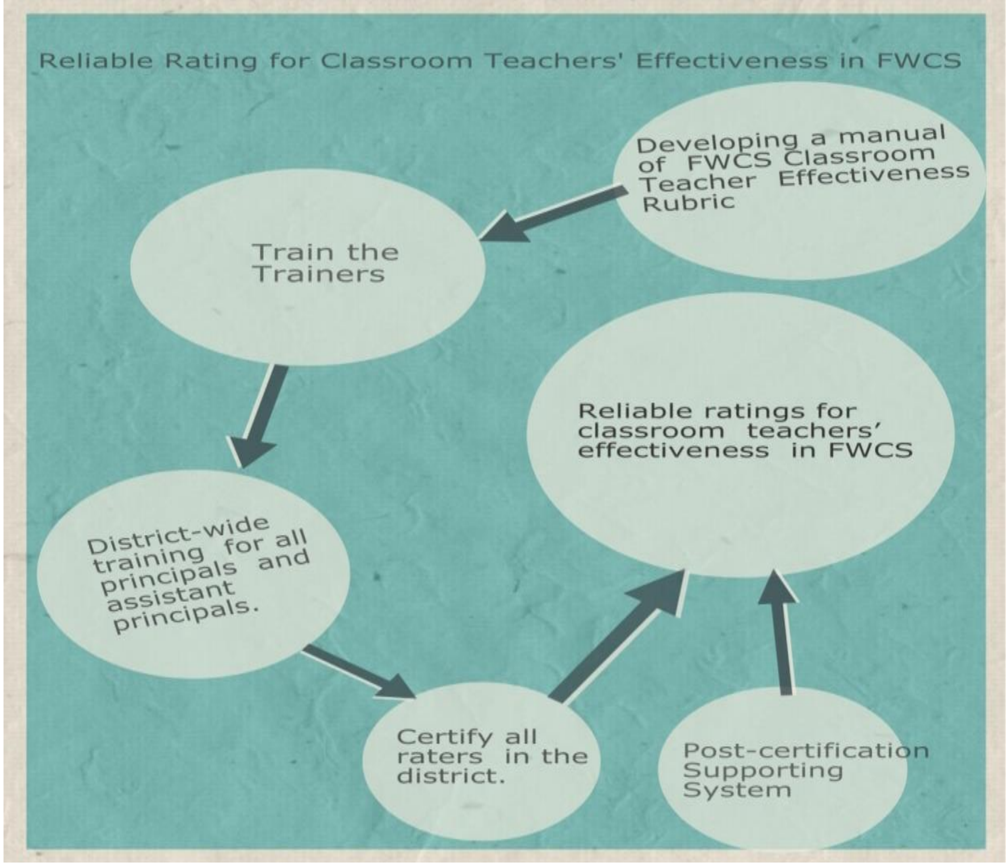


Figure 4. Framework of FWCS Inter-Rater Reliability Training Program


Reliable Rating for Classroom Teachers' Effectiveness in FWCS
Developing a manual of FWCS Classroom Teacher Effectiveness Rubric
Train the Trainers
Reliable ratings for classroom teachers' effectiveness in FWCS
District-wide training for all principals and assistant principals.
Certify all raters in the district.
Post-certification Supporting System

Case Study: Urban Teacher Center

Agency Profile

Urban Teacher Center (UTC), a residency-based teacher preparation program, launched as a non-profit organization in September 2009. In its first year, UTC opened with 39 residents in Baltimore City Schools and charter management organizations in Washington, DC. In 2014–15 school year, UTC welcomed 112 residents for its fifth and largest class yet. In total, 307 residents and teachers serve in 86 schools across Baltimore City and Washington, DC. UTC's inaugural class was prepared in general and special education for literacy and math content areas in grades prekindergarten through nine, leading to dual general and special education licensure for all participants who successfully pass through UTC's performance standards. UTC has since launched a secondary math and secondary English degree and license.

UTC's approach to preparing effective teachers is multifaceted, combining 1) a selective admissions process, 2) intensive training and support through rigorous coursework that is aligned to the teaching practices we expect participants to develop, extensive clinical experience, and ongoing coaching and feedback, and 3) continual evaluation of performance. In the first fourteen months of our four-year program, participants work in classrooms alongside host teachers and receiving on-site coaching. At the same time, they take clinically-based graduate-level courses that introduce them to specific teaching practices and provide immediate opportunities to try those practices with students. Following the residency year, they become teachers of record in urban classrooms where they continue to receive coaching support, and they complete coursework for their master's degree in education.

Policy/Research Questions

UTC seeks to assess and improve the reliability of our rating system. Many of the limitations of teacher practice rubrics can be mitigated by investing in training observers and conducting appropriate oversight to ensure fidelity of their use. In the MET project, raters underwent 17 to 25 hours of training and were required to rate a number of pre-scored videos and achieve a minimum level of agreement with the expert scores prior to certification. MET also monitored rater accuracy on an ongoing basis, and those who failed calibration exercises could not score videos that day.

Recognizing the need to ensure consistent application of teacher rubrics, UTC has adapted the practices used in the MET project. To support consistency in communication to participants regarding what effective teaching looks like, we train and calibrate coaches on the Teacher Practice Rubric before they begin classroom observations. UTC's Curriculum and Professional Development team holds faculty institutes at the start of each semester, during which coaches participate in norming activities. Coaches also complete calibration exercises to assess inter-rater reliability. Following these exercises, we generate individualized reports to share with coaches; these reports compare the coach rating to the master rating for each indicator. We ask coaches to review descriptors of any indicators when the coaches' ratings are not aligned to the master rating.

Even when observers are well trained, a single observation conducted by a single observer is a fairly unreliable estimate of a teacher's practice (Ho & Kane, 2013). In an effort to identify ways to enhance reliability, **UTC focused on two questions: 1) can paired observations increase inter-rater reliability, and 2) how does the number of items and observations influence reliability of classroom observation scores?** Enhancing inter-rater reliability is intended to ensure consistency in the feedback provided to UTC's participants regarding the quality of their teaching practice in specific areas, while enhancing reliability of the overall score is critical to support the use of these scores in making high-stakes decisions regarding whether participants continue in the program.

Project Scope and Timeline

Paired observations took place throughout fall of 2014. UTC asked a lead clinical faculty member to work with new coaches and with coaches whose inter-rater reliability was low based on the calibration exercises conducted in August 2014. During a paired observation (typically about 45 minutes), the lead clinical faculty member and the teacher's coach observed the teacher's lesson together, taking notes as they observed. Immediately following the observation, the lead clinical faculty member and the coach independently rated various indicators of teaching practice based on their interpretation of the evidence. After completing independent ratings, the pair was asked to compare their ratings and resolve discrepancies through discussion. As part of this process, the SDP Fellow attended several of the paired observation sessions and took notes on how the pairs resolved discrepancies. In addition, the SDP Fellow compared inter-rater reliability from the calibration exercises that took place in August 2014 to those that occurred in January 2015.

To address the second research question, in March of 2015, the SDP Fellow conducted a generalizability study to assess multiple elements of the observational system, which inter-rater agreement measures cannot do (Hill, Charambolous, & Kraft, 2012). This study provides empirical evidence regarding how the number of items rated and the number of lessons observed affected reliability of the score. The SDP Fellow constructed a data file that contained ratings on all observations that took place during the 2013–14 school year and generated a subset of observations to analyze. The analytic sample consisted of first-year teachers for whom we had at least three observations in which all 19 indicators were rated. After establishing the proportion of variance attributable to participants and items, the variance components were used in the decision study to shed light on how the number of items and observations affected reliability of the overall score.

Results/Impact

Trends in Inter-Rater Reliability: Figure 5 displays the inter-rater reliability of the nine coaches that participated in the paired observations. The blue bar represents the percent of exact agreement based on the average of three inter-rater reliability exercises at the August Institute, while the gray bar represents the raters' exact agreement with the master rating based on the January 2015 inter-rater reliability exercise. Of the nine coaches who participated in paired observations, seven had a higher proportion of exact agreement with the master rating in the January inter-rater reliability exercise compared to their exact agreement in the August exercises. Coach C remained about the same, and Coach I performed somewhat worse in the January exercise compared to the August exercises.

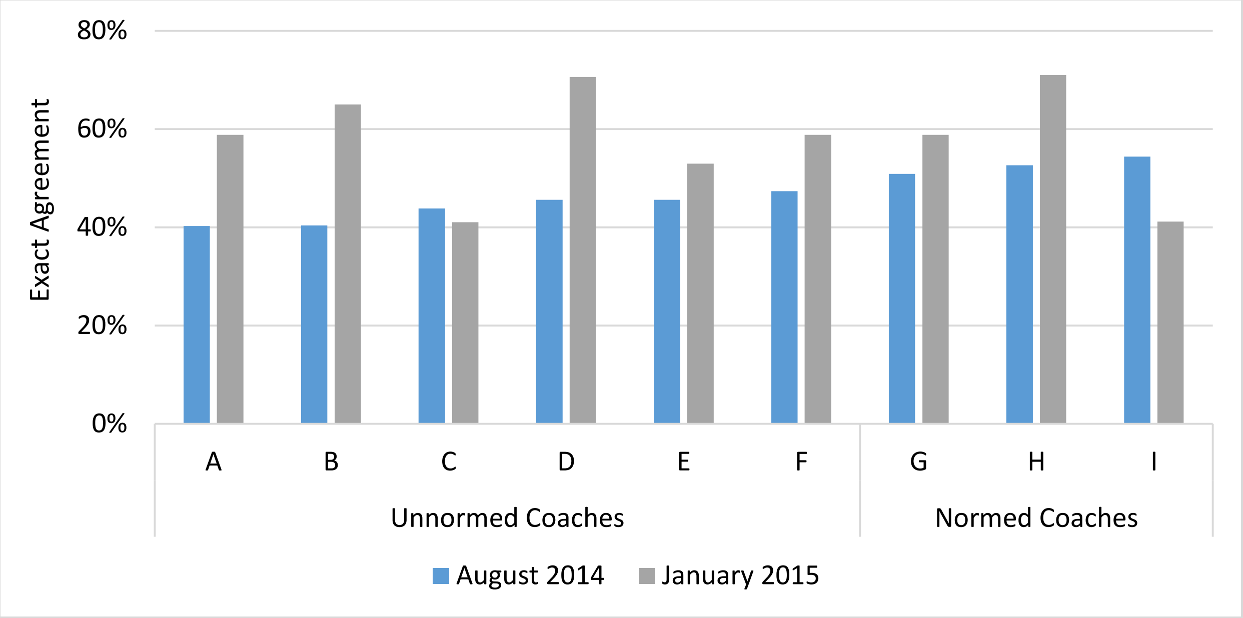


Figure 5. Exact Agreement Between Coaches and Master Ratings, August 2014 and January 2015.

Across the nine coaches, inter-rater reliability in January averaged about 11 percentage points higher than inter-rater reliability in August. The six unnormed coaches averaged an increase of 14 percentage points, from 44% exact agreement in August to 58% in January. Though pairing did not result in universal improvement in inter-rater reliability, the process was highly regarded by the coaches participating in the pilot, who viewed the pairings as an effective form of professional development and appreciated the opportunity to have a one-on-one discussion with lead clinical faculty regarding how to interpret the evidence in light of the descriptions in the Teacher Practice Rubric.

There are a few caveats regarding the results of the pair observation intervention. First, this is an extremely small sample. Second, because we did not assign a control group of unnormed coaches to a "no treatment" condition, we cannot be certain that those coaches improved due to the paired observations. It may be that unnormed coaches benefit more from the paired observations than normed coaches, but it is possible that unnormed coaches would have improved considerably even without this intervention. Finally, paired observations are time-consuming with what might be considered a high opportunity cost. In piloting the program, UTC strategically selected coaches who either had low exact agreement during the training sessions or were recent hires. Although there is interest in expanding the paired observations to all coaches, UTC may need to continue to target specific coaches, since this is a resource-intensive approach to enhancing inter-rater reliability.

The Impact of Number of Items and Observations on Reliability: The SDP Fellow placed at UTC conducted a generalizability study to decompose variability teachers' observation scores into meaningful components of variance. This was followed by a decision study to determine how reliability is affected by the number of items in the Teacher Practice Rubric and the number of observations conducted. The decision study was also to identify scoring protocols that will maximize precision for minimal cost. These analyses are based on a subset of the observations conducted during 201314 school year. Specifically, we used data from 40 teachers with at least three observations that included scores on all 19 indicators. One limitation of this study is that we cannot partition out the variance that is due to rater, because participants were observed by their instructional coaches and in most cases, the coach remained the same throughout the year.

The results of the decision study indicated that adding observations and rubric categories produces diminishing returns to reliability. Reliability improves most markedly (from .70 to .81 at 10 items) when increasing from 1 to 2 observations. Adding a 3rd observation increases reliability to .85, and a 4th results in .87. Similarly, when we increase the number of items from 2 to 4 (at 2 observations), reliability increases from .63 to .73, but reliability only increases from .79 to .81 when we increase the number of items from 8 to 10.

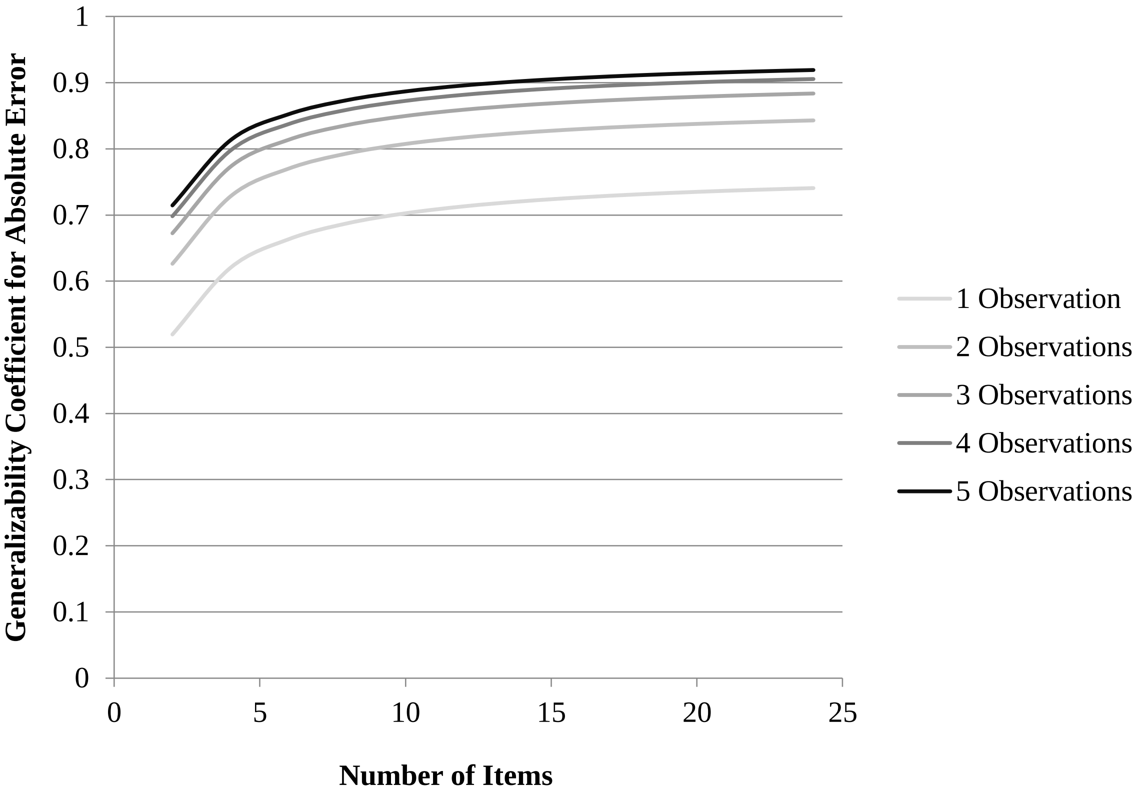


Figure 6. Generalizability Coefficient for Absolute Error.

Because we see a relatively big increase in reliability when we move from one to two observations, UTC plans to use an average of two observations for each summative evaluation of our participants. While the Teacher Practice Rubric is undergoing revisions, we do not plan to reduce the number of indicators, as we are continuing to explore which indicators are most strongly related to student gains and future effectiveness.

Case Study: Oklahoma State Department of Education

Agency Profile

The Oklahoma State Department of Education (OKSDE) determines education policies and directs the administration and supervision of the public school system of Oklahoma. As of the 20142015 school year, the agency serves 688,300 students in 1,795 schools in 543 districts and employs 40,227 full-time equivalent (FTE) teachers. The student population is 50.8% White, 15.6% Hispanic, 14.6% American Indian, 9.1% Black, 7.7% Multi-racial, and 2.2% Asian or Pacific Islander.

One OKSDE priority is the development and implementation of a Teacher and Leader Effectiveness (TLE) evaluation system to inform instruction, create professional development opportunities, and improve teaching. The multiple-measures evaluation system consists of both qualitative teacher evaluations and quantitative student growth measures. For teachers in tested grades and subjects, student growth is calculated using a value-added model (VAM).

Policy/Research Questions

In designing Oklahoma's VAM, OKSDE had to choose what student and school characteristics and what prior year tests to include. Each decision involved specific trade-offs that impacted actual and perceived fairness, understandability, precision and comparability of the results. Therefore, OKSDE conducted research to explore the trade-offs of the chosen specification and to understand how the resulting value-added scores related to observable school- and teacher-level characteristics. The research addressed the following questions:

1. What are the advantages and disadvantages of Oklahoma's VAM in terms of fairness, accuracy, and usefulness?
2. What is the relationship between VAM and school-level characteristics such as the percent of students in poverty and other demographic factors?
3. What is the relationship between VAM and teacher-level characteristics such as subject taught or years of teaching experience?

Project Scope and Timeline

In fall 2013, OKSDE held meetings with VAM experts and Oklahoma educators to make choices regarding VAM specification. Table 1 summarizes the characteristics of OKSDE's VAM that resulted from this process. In summer 2014, results of the first year of VAM were available and this analysis was completed.

Table 1. Key Charactieristics of Oksde's Value-Added Model (VAM)

Characteristic	OKSDE Model
Developer	Mathematica
Model	Two-stage linear regression model
Tests Used	Math: OCCT math in grades 4 through 8, algebra I in grades 8 and 9, algebra II in grades 9 through 11, and geometry in grades 9 through 12; Reading: OCCT reading in grades 4 through 8 and English III in grade 11
Student Characteristics Included	Poverty status, gender, race/ethnicity, existence of an individualized education plan, limited English language proficiency status, transfers between schools during the school year, and prior year school attendance
Teacher Characteristics included	None
School Characteristics Included	None
Minimum Number of Students	10
Adjustments for Measurement Error	Shrinkage
Reasons for Student Exclusion	(1) Conflicting post-test scores (2) Missing pre-test score from same content area (3) Skipped or repeated a grade (4) Not linked to an eligible teacher

Results/Impact

Advantages and Disadvantages of Oklahoma's VAM: Oklahoma's VAM does not include school-level characteristics. The advantage of this choice is that Oklahoma can compare teachers across the entire state. If Oklahoma were to include school-level variables, the state would only be able to compare teachers within schools and would make the implicit assumption that the average teacher skill at all schools is the same, which is arguably not the case.

At the same time, the choice to exclude school-level variables does come with several limitations. By not including school factors, it is impossible to distinguish between student growth attributed to the teacher and growth attributed to the school. It would be impossible, for instance, to distinguish between the impact of a teacher at a certain school and the supports, programs and climate at that school, all of which may also contribute to a student's growth. As a result, a teacher at a school with more effective supports, programs, and climate may be more likely to receive a higher value-added score. This is because the effect of these school-level characteristics will be included in his or her value-added score. Therefore, Oklahoma's VAM may systematically favor teachers at certain schools and disadvantage teachers in others.

Finally, Oklahoma's model includes scores from only one prior year. This has the advantage of including more students in the model as mobility, missing data, and other factors make it difficult to link students to their prior test scores for two consecutive years. It also allows scores to be calculated starting in fourth grade rather than fifth grade. At the same time, the inclusion of only one prior year of data makes Oklahoma's VAM more sensitive to the performance of a student in a single prior year than models based on two years.

School-Level Characteristics: As Figures 7–10 demonstrate, school-level value-added scores are significantly correlated with school-level characteristics such as the percent of students in poverty, the percent of students on individualized education plans (IEPs), the percent of English language learning (ELL) students, and the percent of minority students. In all cases, larger percentages of students in these categories were associated with lower school-level value-added scores, on average. There are two main explanations for this.

First, the previously discussed exclusion of school-level factors makes it impossible to separate the effect of schools and individual teachers at the schools on student growth. If there is a relationship between the existence of school-level supports, programs and learning environments, and school-level demographic factors such as poverty or the percent of students on IEPs, this effect could be explained by the exclusion of school-level factors. As an example, if students in low-poverty schools tend to have supports, programs, and learning environments associated with higher student growth relative to their peers in high-poverty schools, we would expect school-level value-added to be negatively related to the percent of students in poverty. In other words, as average poverty increased, we would expect school-level value-added to decrease.

Another explanation for these results is real differences in teaching ability at schools with different student populations. Teachers are non-randomly distributed across schools, and more effective teachers may sort into schools with fewer minority students or students in poverty. Schools with more disadvantaged populations, moreover, also have a disproportionate numbers of new teachers, which evidence shows are less effective, on average, also contributing to this result.

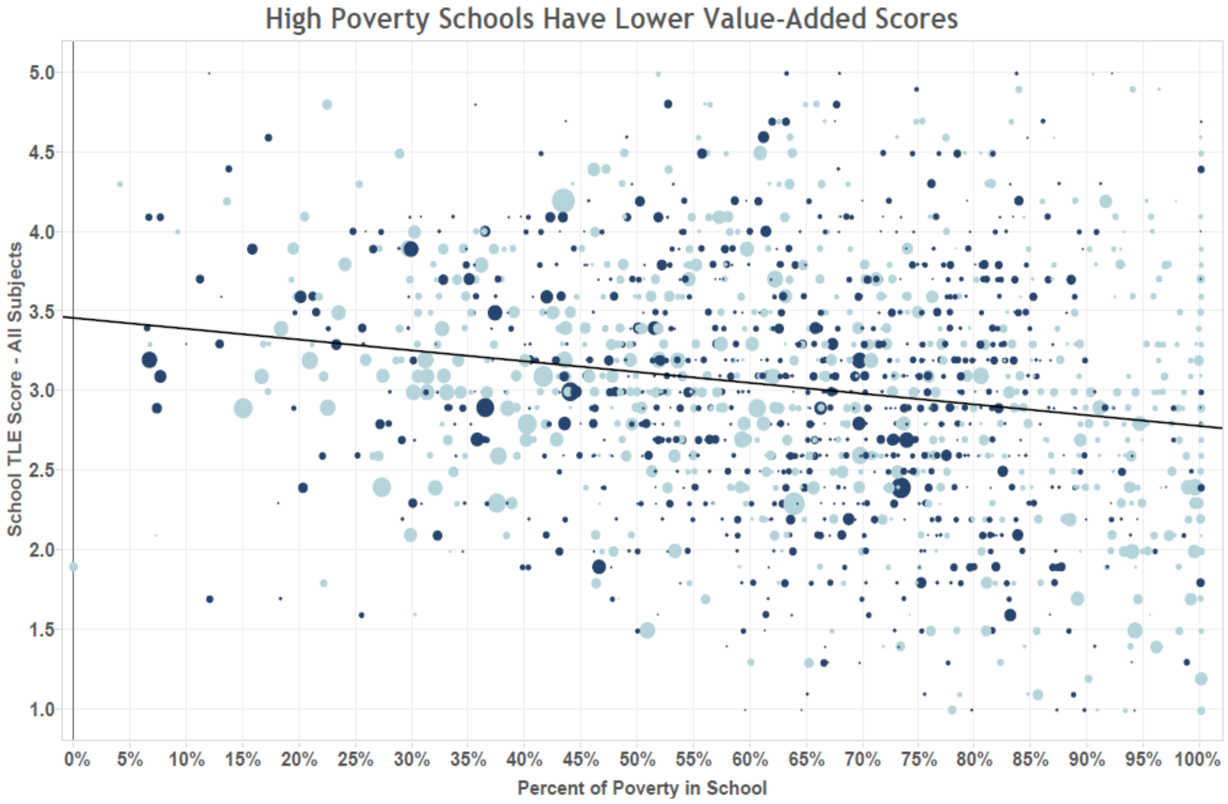


Figure 7. School-Level Poverty and Value-Added Scores

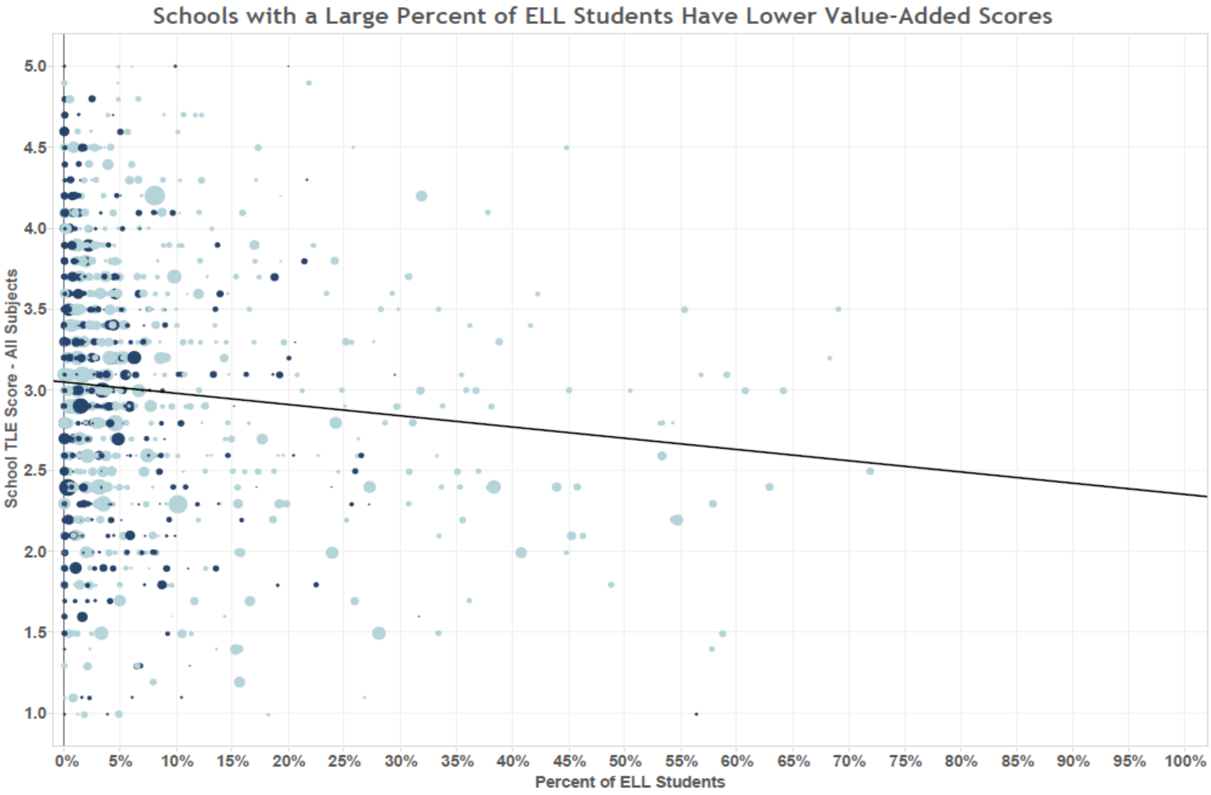


Figure 8. The Percent of Ell Students and Value-Added Scores

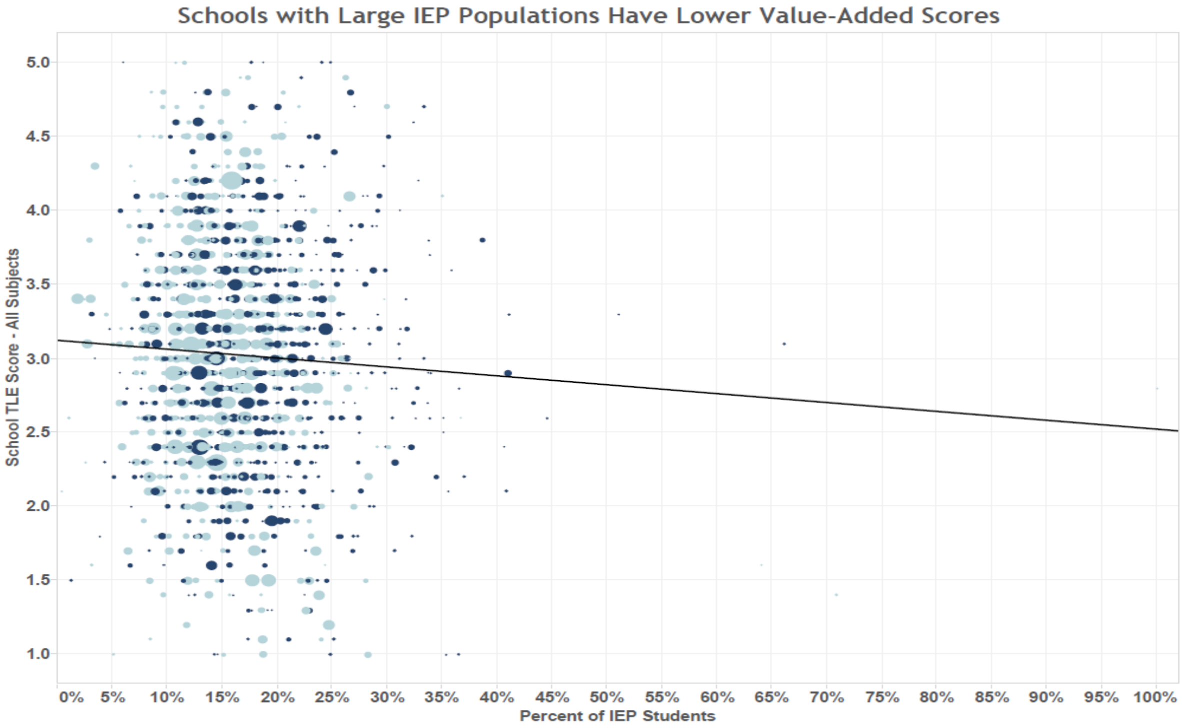


Figure 9. The Percent of Students on IEPs and Value-Added Scores

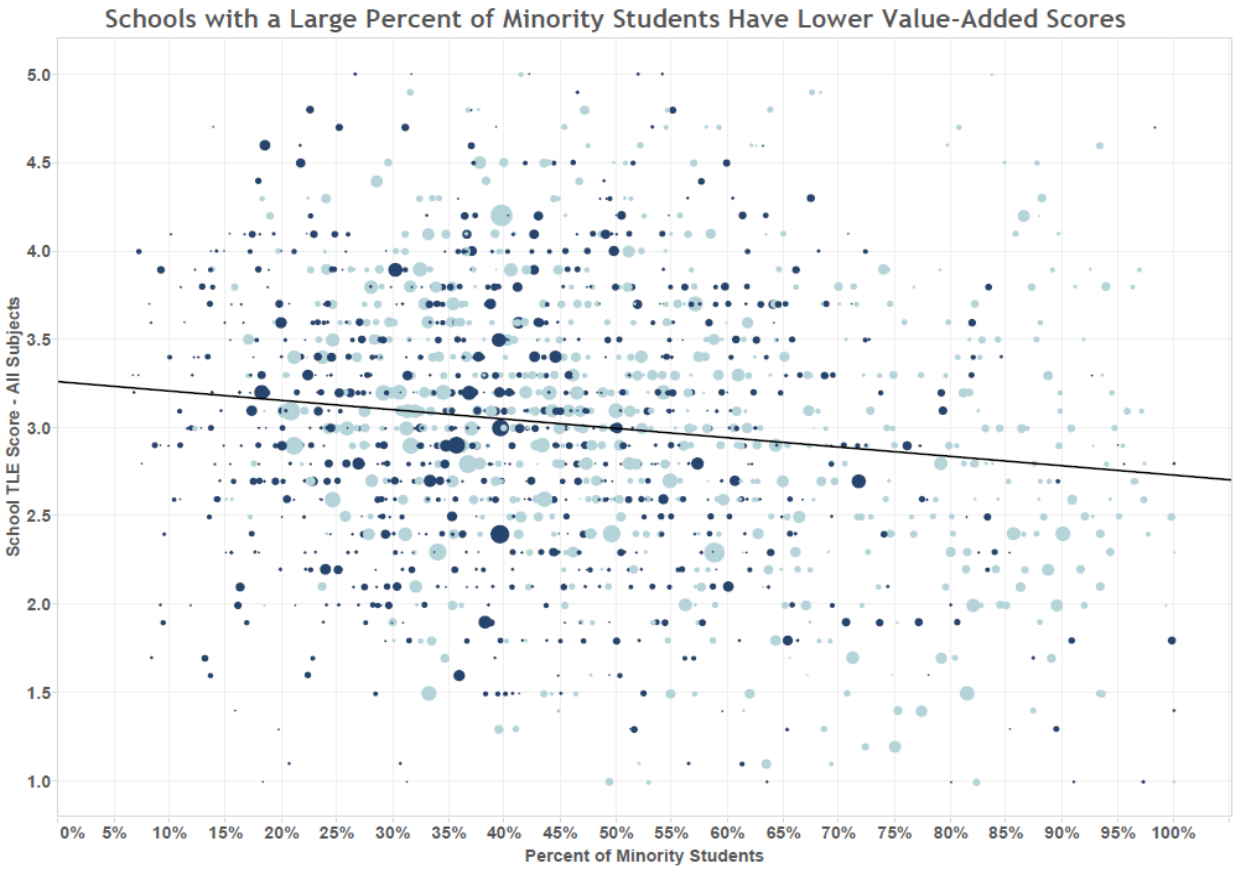


Figure 10. The Percent of Minority Student and Value-Added Scores

Teacher-Level Characteristics: The data revealed several insights on the link between teacher experience and teacher effectiveness. Using administrative data on teacher personnel, the SDP Fellow placed at OKSDE compared years of teaching experience to value-added scores. The analysis included 7,034 unique teachers. Included teachers had between 0 and 48 years of teaching experience with an average of 9.72 years.

As Figures 11 demonstrates, years of experience were positively and significantly related to VAM scores. In other words, teachers with more experience had higher VAM scores, on average. The most significant effect was for first-year teachers, although teachers in later years did demonstrate gains as well. These results also vary by subject. As Table 2 demonstrates, the highest gains to experience for first-year teachers were for OCCT Reading, OCCT Math and Algebra 1 teachers, respectively.

Table 2. Average TLE Score by Subject and Experience Level

Experience	Teachers (n)	All Subjs	OCCT Math	OCCT Read	Algebra 1	Geometry	Algebra 2	ELA3
New	479	2. 62	2. 62	2. 53	2. 67	2. 78	2. 73	2. 77
1–3 years	574	2. 96	3. 03	2. 94	2. 81	3. 06	3. 10	2. 77
4–10 years	1777	3. 00	3. 00	3. 00	3. 04	3. 08	3. 14	2. 96
Greater than 10 years	4204	3. 06	3. 07	3. 06	3. 07	3. 08	3. 02	3. 10
All	7034	3. 01	3. 02	3. 00	3. 02	3. 06	3. 04	3. 01

Interestingly, several experienced teachers did not do better than the average first-year teacher. As Table 3 shows, 33% of all teachers with at least one year of experience had lower TLE scores than the average new teacher. This has significant implications for the teaching workforce. If low-performing, experienced teachers were replaced with new teachers, for example, student achievement would be expected to increase, on average.

Table 3. Percent of Teachers Scoring Worse than First-Year Teachers

Experience	All	OCCT Math	OCCT Read	Algebra 1	Geometry	Algebra 2	ELA 3
1–3 years	40%	36%	35%	52%	39%	33%	30%
4–10 years	34%	35%	31%	33%	34%	30%	31%
Greater than 10 years	32%	32%	29%	31%	38%	39%	33%
All	33%	33%	30%	34%	37%	36%	32%

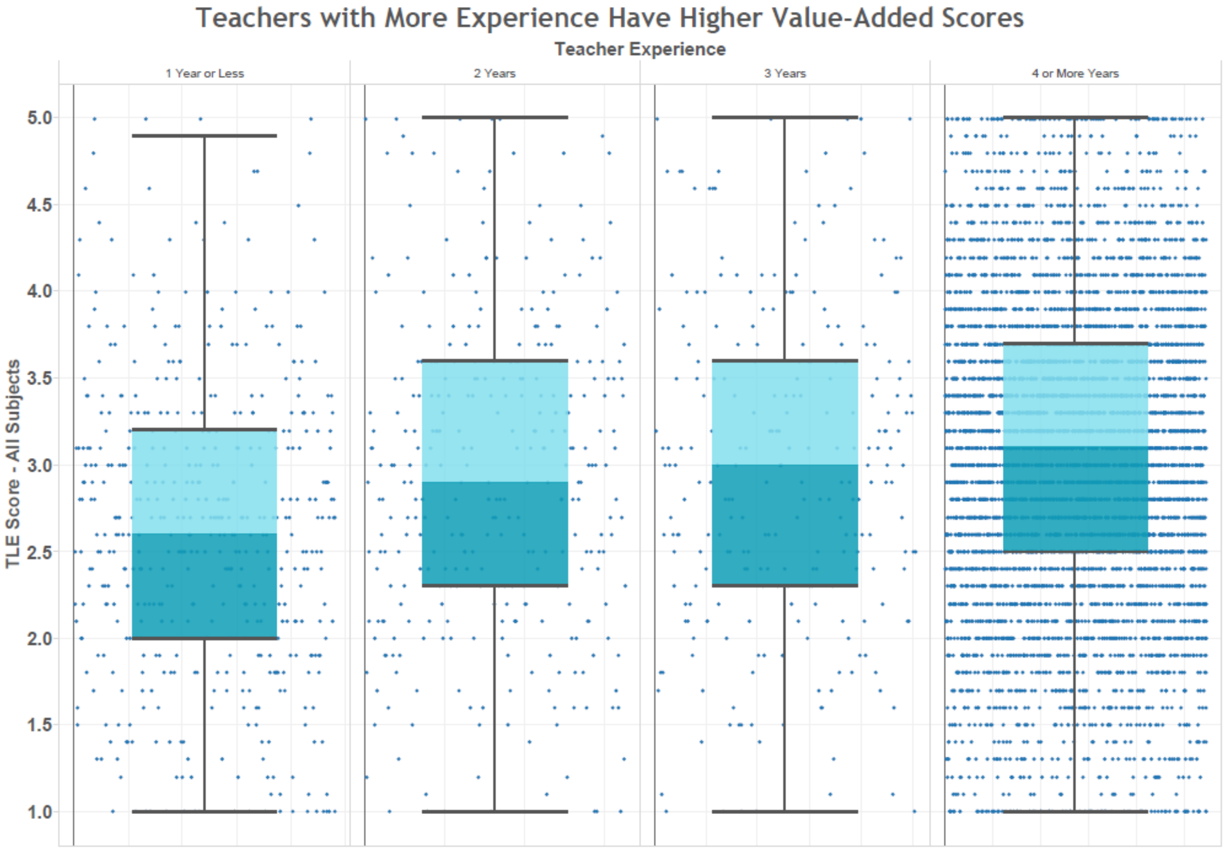


Figure 11. Teacher Experience and VAM Scores

Case Study: Tulsa Public Schools

Agency Profile

Tulsa Public Schools (TPS) is the second largest school district in the state of Oklahoma. Serving approximately 40,000 students with a base of roughly 7,000 employees, TPS is made up of a diverse group of students. The district is represented by approximately 30% Hispanic, 27% White, 26% Black and 6% American Indian students. Within this racially diverse group, nearly 80% of students receive free or reduced-price lunch. Consistent with other large urban districts, a large number of students are identified as English language learners (18. 5%) or on individualized education programs (16. 5%).

In 2010, TPS stated its mission is to provide quality learning experiences for every student, every day, without exception. To accomplish this mission, Tulsa Public Schools identified Teacher and Leadership Effectiveness (TLE) as a key focus area in its 2010 strategic plan. Over the course of the past five years, TPS has designed, developed, and rolled out an observation rubric and evaluation protocol used by nearly 500 school districts in the state of Oklahoma. At the same time, a value-added model has been put in place that now produces a VA estimate for nearly 50% of the 2,400 classroom teachers in Tulsa. Over the past three years, Tulsa has added student surveys to its measures of teacher effectiveness. All of this work has been with the goal of implementing a multiple measures evaluation system in the district.

Policy/Research Questions

Tulsa Public Schools (TPS) aims to create a measurement system that uses rigorous tools to assess teacher effectiveness combined with policy that decreases the probability of a teacher whose true effectiveness is effective or better being exited based on a performance rating that is reported as less than effective.

After selecting the measures that will be used in a multiple measures evaluation system, State Education Agencies (SEAs) and Local Education Agencies (LEAs) must decide how the measures will be combined to create a composite or summative rating of teacher effectiveness. The empirical evidence related to how much weight should be placed on each measure is minimal. In the MET Project policy brief, *Ensuring Fair and Reliable Measures of Effective Teaching,* Cantrell and Kane discuss the tradeoffs between reliability and predictive power when selecting weights in a multiple measures evaluation system. In their brief, Cantrell and Kane identify evidence that suggests value-added should account for 33%– 50% of the weight in a composite score with principal observation ratings making up no less than half of the remaining weight (Cantrell & Kane 2013). Understanding the tradeoff between reliability and predictive power, TPS aims to implement a system that is both reliable and predictive of a teacher’s impact on student performance. In collaboration with the local teacher union, Tulsa Classroom Teachers Association (TCTA), TPS also hopes that the local evaluation system respects the multifaceted nature of teaching – understanding that no measurement tool can perfectly capture the true impact of a teacher.

TPS sought to understand how an innovative evaluation system can 1) decrease the rate of misidentification of effective teachers as less than effective, 2) increase the usefulness of the information for both principals and teachers, and 3) improve overall teacher effectiveness through identification of strengths and opportunities for growth for individual teachers.

Project Scope and Timeline

In the winter of 2013–14, Tulsa Public Schools set out to develop a multiple measures system that used the three measurement tools that were already in place in TPS. [2] The goal was to have a system rolled out and in place for the 2014–15 school year with teachers receiving multiple measures reports at the end of the 2014–15 school year. The district worked closely with the Tulsa Classroom Teacher's Association, the representatives of teachers in Tulsa, to develop a multiple measures evaluation system that was both rigorous and fair. Together six key principles were agreed upon:

1. **Application:** the system would pertain only to classroom teachers (not support staff).
2. **Value:** only high-value, high-quality quantitative measures would be used.
3. **Simplicity:** the system must be easy to communicate and understand.
4. **Scaling:** a consistent scale would be developed for all three measures to avoid misconceptions and alarm. [3]
5. **Equity:** the system would maximize fairness and avoid unintended consequences.
6. **Exiting:** mutually-agreeable protocols would be established to determine how the new evaluation system would be related to exiting decisions.

Using these six principles as a basis for the development of a multiple measures evaluation system, design work and analysis of these proposed designs began in April 2014.

To understand the impact of the proposed multiple measures systems on the distribution of teacher effectiveness, historical data from the prior three school years was used to estimate the percent of teachers who would fall into each classification of teacher effectiveness. [4] Under the historical system where principal observations acted as the sole for-

[2] The teacher evaluation system and value-added system were designed and implemented beginning in 2010. Student surveys were implemented in 2013 as a pilot and district-wide implementation of surveys occurred in the 2014–15 school year.

[3] At the time of development, teacher observation scores were reported on a 1–5 scale, value-added scores were reported on a 0–5 scale and student surveys were reported on a 0%–100% scale.

[4] The state-defined classification system is made up of 5 ratings based on a continuous scale from 1–5.

stakes measure of teacher effectiveness, zero teachers had been identified as ineffective and fewer than 2% of teachers had received a rating of needs improvement over the course of the two prior school years. This trend of identifying very few teachers as less than effective is consistent with other large school districts across the nation (Weisberg et al. 2009). TPS leaders hoped that the resulting multiple measures system would help to identify low performing teachers who had been receiving qualitative ratings of effective or better while also helping principals to identify teachers who may need additional support based on the quantitative ratings of teacher effectiveness.

Internal analyses of historical data indicated that the traditional weighting approach for multiple measures used by many states and districts would result in 13%–18% of teachers being rated as less than effective. This staggering increase was not acceptable to the local teachers association and was not practical for the district given the number of mandatory exits it would prompt. In addition to the practical and political concerns generated by the weighting system, the internal researchers assessing the impact of the proposed multiple measures system worried that a system that assigned rigid weights to the three measures would result in the misidentification of many teachers as less than effective. Together the researchers worked alongside district leaders and the teachers association to develop a system that met the six principles outlined above and respected the imperfect nature of the qualitative and quantitative measures that would make up the multiple measures system.

Results/Impact

The Resulting System: The TPS multiple measures evaluation system does not produce a single summative score; instead, all three scores (observation, student surveys, value added) are reported independently in a single report (see appendix 44 for a sample multiple measures report). In addition to not reporting a single summative score, another primary departure in the Tulsa system when compared to the weighted system used in many states and districts is the rescaling of the quantitative measures in Tulsa’s system. Tulsa’s system reports categorical ratings of below average, average, and above average for value-added and student

Ineffective (<1. 80), Needs Improvement (1.80 – 2.79), Effective (2.80 – 3.79), Highly Effective (3.80 – 4.79), Superior (4.80 – 5.00).

surveys. These ratings are based on a teacher's score relative to the district mean where below average represents performance more than one standard deviation *below* the district mean, average represents performance within one standard deviation of the mean and above average represents performance more than one standard deviation *above* the mean.

Table 4.

		Student Surveys			
		Below Average	Average	Above Average	No Data
Value Added	Below Average	7	38	17	9
	Average	46	594	200	93
	Above Average	1	65	36	6
	No Data	98	185	601	451

The multiple measures report serves three primary purposes:

1. Provide teachers with a summary of their performance on all three measures of effectiveness in a single location.
2. Provide principals with a document to analyze and for discussions with teachers about their performance and professional goals based on the data.
3. Provide school and district administrators with a framework for exiting ineffective teachers.

First Year Implementation of the System: The system described above was rolled out to teachers in the 2014–15 school year across Tulsa Public Schools. As a lead up to the implementation and rollout of the new evaluation and reporting protocol, district administrators from the office of Teacher and Leader Effectiveness presented in person to each of the 90+ school sites over the course of two months. Draft reports were developed in collaboration with teachers and principals beginning in November and continuing through early

March. Final Multiple Measures Reports were made available to teachers and administrators in the final weeks of the school year.

Next Steps: The transition to an evaluation system that includes multiple measures of teacher effectiveness is both a challenge and opportunity that many districts and states will likely face in the coming years. Whether implementing an evaluation system with multiple measures for the first time or revamping a current evaluation system, there are multiple considerations to take into account. There will inevitably be tradeoffs with the decisions that are made. These decisions and tradeoffs should not be made in a vacuum; instead, these decisions should be made in collaboration with a diverse group of stakeholders including, but not limited to, district leadership, principals, teachers and policymakers. The collaborative nature of the development process in Tulsa helped to ensure that all stakeholder groups understood the purpose of and were bought into the new evaluation process.

Further research is necessary on the impact of multiple measures systems on human capital and professional development processes. TPS will continue to monitor and evaluate the effects of its multiple measures system on the retention, development, and placement of teachers within the district. It is critical that TPS and other districts that are early adopters in the multiple measures evaluation space remain flexible and willing to adjust as additional research suggests the adoption of new practices.

Lessons Learned

Although each agency's case study focuses on a different aspect of teacher evaluation systems, they share a common goal of promoting reliable and valid measures of teacher effectiveness and ultimately raise student achievement. When considered together, several key themes emerge from the case studies that may benefit other organizations attempting to develop or refine their approach to teacher evaluation.

First, the communication, training, and support required to implement a teacher evaluation system should not be overlooked. For FWCS, initial training was spotty at best. By the time the need arose to tackle inter-rater reliability, few of the initially trained administrators were still in the district. Initial training of evaluators is an essential first step, but is insufficient to ensure that teacher evaluation systems will support achieving intended goals in the long run. The calibration process is useful for quality control purposes. However, in districts like FWCS and organizations like UTC, replacing observers who perform poorly on calibration tests is not always a desirable option. At FWCS, for example, launching a process modeled after the MET Project produced a principal leadership team excited about their own professional learning and determined to share the calibration process with their peers. Another approach is to use data from the calibration process to target follow-up interventions to the observers most in need of additional support. UTC's follow-up intervention, the paired observations, yielded promising results in improved inter-rater reliability. Agencies committed to the process of increasing reliability and validity of evaluative measures should expect a long-term commitment of time and resources to enact the evaluation processes, sustain support from relevant stakeholders, and provide on-going guidance to institutionalize the processes.

Second, policymakers encounter numerous decisions when designing teacher evaluation systems, and these decisions involve making trade-offs. It is important that policymakers consider the impact of these decisions on the validity and reliability of teacher ratings. For example, while a greater number of observations will almost certainly increase reliability, each additional observation incurs costs. As another example, including school-level fixed effects in a VAM may be viewed as more valid or fair as it accounts for the impact of the school on student achievement, but that decision would mean that teachers are only compared to others within the same school, not across an entire SEA or LEA. Understanding these trade-offs will help policymakers make choices that best reflect the values and goals of their agency.

Third, in developing the evaluation system, policymakers will also need to consider how their decisions impact stakeholders. An evaluation system that is designed in a manner that is perceived as punitive by educators is unlikely to achieve the ultimate goal of improving teaching quality and, subsequently, student achievement. In addition to stakeholder perception, practical realities must be considered when designing evaluation policies. At TPS, analyses revealed that applying the same fixed weights for multiple measures used by a similar school district in a different state would result in nearly 20% of teachers falling below the state defined cutoff for effective. This outcome would be not only unpopular but potentially over identify the number of teachers in need of support thereby not allowing for focused interventions for actual struggling teachers. As SDP Fellows, an important aspect of our work is to clearly present evidence that is brought to bear on decision-making, and such presentations can be a valuable tool as part of a process for gathering feedback and garnering support from stakeholders.

Finally, we must be humble about the precision and reliability of the measurement tools we have available. Even in conservative observation frameworks, value added models and multiple measures systems, some teachers will be misclassified—meaning some effective teachers will receive ratings of less than effective while some ineffective teachers will receive ratings of better than effective. Organizations that approach the evaluation process with an understanding that the measures used to evaluate teachers are not perfect will be well suited to use these tools to accomplish the end goal of improving the effectiveness of teachers in their organization.

To better understand the advantages, disadvantages, and trade-offs of decisions regarding the overall evaluation model and its components, and to select the best model for an agency's purposes, research on alternative specifications of the model during both the pilot phase and an on-going basis is important. In addition, future research could shed light on how to best use the information from teacher evaluation systems to generate high-quality, targeted, and efficient professional development.

References

American Statistical Association (ASA). (2014, April). *ASA Statement on Using Value-Added Models for Educational Assessment.* Alexandria, VA: Author. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Bill & Melinda Gates Foundation. (2011). *Learning about teaching: Initial findings from the Measures of Effective Teaching project.* Bellevue, WA: Author. Retrieved from www.gatesfoundation.org/college-ready-education/Documents/preliminary-findings-research-paper.pdf

Balch, R. T. (2012). *The validation of a student survey on teacher practice(Doctoral dissertation,* Vanderbilt University). Retrieved from http://mystudentsurvey.com/wp-content/uploads/2012/06/Balch-Student-Surveys-2012.pdf

Corcoran, S., & Goldhaber, D. (2013). Value Added and It's Uses: Where You Stand Depends on Where You Sit. *Education Finance and Policy, 8*(3), 418–434.

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267–297. doi: 10. 1002/pam. 21818

Goldhaber, D., & Loeb, S. (2013). What Do We Know About the Tradeoffs Associated with Teacher Misclassification in High Stakes Personnel Decisions?. *The Carnegie Knowledge Network*.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Education Assessment, Evaluation and Accountability, 26*, 5–28.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher, 41*(2), 56–64.

Ho, A. D., & Kane, T. J. (2013). *The Reliability of Classroom Observations by School Personnel.* Research paper. MET Project. Bill & Melinda Gates Foundation, Seattle, WA. Retrieved from http://www.metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf

Jerald, C. (2012). *Ensuring Accurate Feedback from Observations.* MET Project. Bill & Melinda Gates Foundation, Seattle, WA.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review* (2), 130–144.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (2014). *Designing Teacher Evaluation Systems.* San Francisco: Jossey-Bass.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. MET Project Research Paper, Bill & Melinda Gates Foundation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Minner, D., & DeLisi, J. (2012). *Inquiring into Science Instruction Observation Protocol (ISIOP): User's Manual.* Education Development Center.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review, 102*(7), 3184–3213.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking Teacher Evaluation in Chicago Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation.* Chicago, IL: Consortium on Chicago School Research.

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review, 102*(7), 3628–3651.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* The New Teacher Project. New York, NY.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts.* Brown Center on Education Policy at Brookings Institution, Washington, DC.

FWCS Appendix

Project 1: *Analyses of Validity of FWCS's Classroom Observation Rating*

The process involved the following steps:

1. Review district policies and implementation procedures regarding classroom observation rating.
2. Communicate with the district's leadership team about the necessity of the project to achieve permission and support from district decision-makers.
3. Design the study including framing questions, identifying measures and data used, specifying data collection procedures, and planning analytic strategies.
4. Collaborate with Technology department to locate and retrieve data needed.
5. Check data quality and clean the data for analyses.
6. Analyze the data to answer research questions.
7. Interpret the results.
8. Report results and suggest policy implication to leadership team.

The Validity of FWCS's Classroom Observation Rating

The following outlines the one-year milestones:

- January 2014: RFI and RFQ consultant hired for project management support;
- June–July 2014: Fellows complete project review of literature and White Paper;
- September 2014: Cabinet approves the scope of the project;
- October 2014–July 2015: The District extends the contract of Dr. Kay Psencik, a consultant with *Learning Forward*, to support the project and the creation of an inter-rater reliability training manual;
- October 2014: The district purchased rights to the Calibration and Certification Engine from Empirical Education;
- December 2014: Fellows submitted the Theory of Action and research proposal to cabinet who selected a leadership team selected;
- December 2014–June 2014: The leadership team meets in full day sessions to define terms on the Indiana RISE rubric and norm videos. The Leadership Team assumes responsibility for all professional learning surrounding Inter-rater reliability. Which includes: meeting protocols, a summer PL session for principals and assistant principals, and provided one-on-one support for principals struggling with becoming certified;
  January 2015: Leadership team defines *Rigor*. After teacher evaluation shows that teachers had difficulty teaching with rigor. Using a qualitative protocol, the district collected nearly 900 inputs in order to create a definition of Rigor;
  December 2014–June 2015: The district realizes that the CCE lacks enough exemplars of highly effective and effective teachers. The district contracts with School Improvement Network to create 12 videos mirroring the MET Project videos. Leadership team balances the teachers list by race, gender, sexual orientation, and teaching level in order to capture the widest view of the district;

- July 14, 2015: This administrative professional learning opportunity took over one complete day of our normal District Principal Institute. Every assistant principals, administrative intern, and Guidance Coordinator was invited to participate in an overview of inter-rater reliability. While this overview was taking place principals were clustered in “feeder” patterns and took their first certification test.

UTC Appendix

Prior to undertaking the generalizability study to establish the amount of variance in observation scores attributable to different sources, UTC determined the coefficient alpha for each observation. Coefficient alpha is .96 for the first observation, .94 for the second observation, and .97 for the third observation. Thus, the estimated expected correlation between two replications of the measurement procedure, where items like these items are randomly drawn and administered, is between .94 and .97. We estimate that 94% or more of the observed score variance can be accounted for by true score variance.

The correlation of average scores from different observations ranges from .71 to .81. This is the test-retest reliability, an estimate of the reliability when we consider items as fixed and occasions as random. It is an estimate of the expected correlation between two replications across different occasions if items are held constant.

The table below lists the sources of variance, the amount of variance, the standard error, and the percent of total variance explained by each source. The p variance is the estimated “true score” variance, or the “good” variance that distinguishes among teachers. The i variance is the variance of individual item difficulties; it is neutral variance as long as everyone takes the same items, and as long as only relative (not absolute) position on the scale matters.

Source	$\widehat{\sigma}_v^2$	$\widehat{\sigma}_v$	Percent
p	0. 4419	0. 6648	39. 18%
i	0. 1184	0. 3441	10. 49%
pi	0. 1186	0. 3445	10. 52%
o|p	0. 1316	0. 3628	11. 67%
io|p,e	0. 3175	0. 5635	28. 15%

The pi variance is the variance of person-item interactions, and error. The o|p variance component refers to the variance of person-observation interactions, and error. These variances reflect the extent to which particular persons score higher or lower on particular items (or observations) above and beyond what is predicted by person ability and item (or observation) difficulty. Both pi and o|pvariance are undesirable with respect to relative error. The o|p variance is undesirable because the relative position of the teacher’s score would change if they were observed on a more or less favorable occasion. The pi variance is undesirable because if we sampled different items, teacher rankings (relative position) would change. However, because UTC standardizes items across persons (that is, all participants are observed on the same items), the pi variance is neutral with respect to relative error in this case.

It is not necessarily the case that greater percentages indicate more important sources of variance, since the variance components are estimated variances of the distributions of *single* effects. Reported scores are not based on a single item, but averages of items (and possibly, averages of observations). Thus, the results of the decision study (presented in the main text) describe the importance of each source of error in terms of their impact on reliability.

Tulsa Public Schools Appendix

Observation scores:

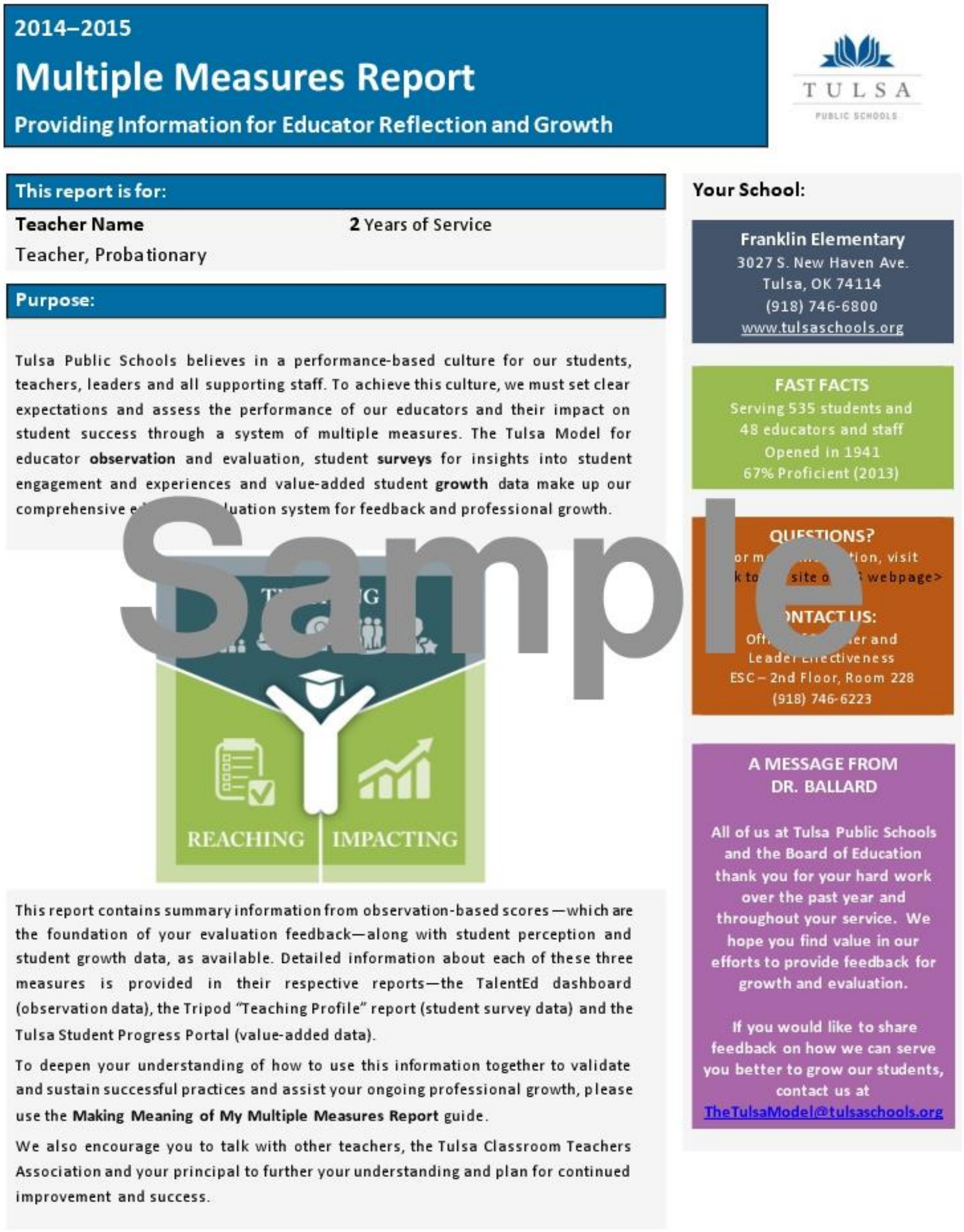
- Reported as a numeric value with a single “combined weighted average” of 1–5.
- The observation score serves as the basis for any personnel action including teacher exits and advancement.
- There are no changes from the way observations have been reported in the past.
- Current year data

Value added scores:

- Reported categorically as “Below Average,” “Average,” or “Above Average”
    - Below Average – overall VA score of <2.00 (greater than one standard deviation *below* the district mean).
    - Average – overall VA score between 2.00 – 4.00 (within one standard deviation of the district mean).
    - Above Average – overall VA score of >4.00 (greater than one standard deviation *above* the district mean).
- Reported as “Data Not Available” for teachers who do not teach in a tested grade and subject.
- Up to 3 year VA score is used. Data is lagging by one school year.

Student Survey Scores:

- Reported categorically as “Below Average,” “Average,” or “Above Average”
    - Below Average – greater than one standard deviation *below* the district mean.
    - Average –within one standard deviation of the district mean.
    - Above Average –greater than one standard deviation *above* the district mean.
- Reported as “Data Not Available” for teachers who do not receive student survey results.
    - These are typically teachers in small classes as you must have 10 students in a single class to participate in the student surveys.
    - Current year data from *better* of two administrations.



2014–2015

Multiple Measures Report

Providing Information for Educator Reflection and Growth

TULSA
PUBLIC SCHOOLS

This report is for:

Teacher Name **2** Years of Service
Teacher, Probationary

Purpose:

Tulsa Public Schools believes in a performance-based culture for our students, teachers, leaders and all supporting staff. To achieve this culture, we must set clear expectations and assess the performance of our educators and their impact on student success through a system of multiple measures. The Tulsa Model for educator **observation** and evaluation, student **surveys** for insights into student engagement and experiences and value-added student **growth** data make up our comprehensive educator evaluation system for feedback and professional growth.

This report contains summary information from observation-based scores—which are the foundation of your evaluation feedback—along with student perception and student growth data, as available. Detailed information about each of these three measures is provided in their respective reports—the TalentEd dashboard (observation data), the Tripod "Teaching Profile" report (student survey data) and the Tulsa Student Progress Portal (value-added data).

To deepen your understanding of how to use this information together to validate and sustain successful practices and assist your ongoing professional growth, please use the **Making Meaning of My Multiple Measures Report** guide.

We also encourage you to talk with other teachers, the Tulsa Classroom Teachers Association and your principal to further your understanding and plan for continued improvement and success.

Your School:

Franklin Elementary
3027 S. New Haven Ave.
Tulsa, OK 74114
(918) 746-6800
www.tulsaschools.org

FAST FACTS
Serving 535 students and
48 educators and staff
Opened in 1941
67% Proficient (2013)

QUESTIONS?
For more information, visit
<link to [illegible] site or [illegible] webpage>

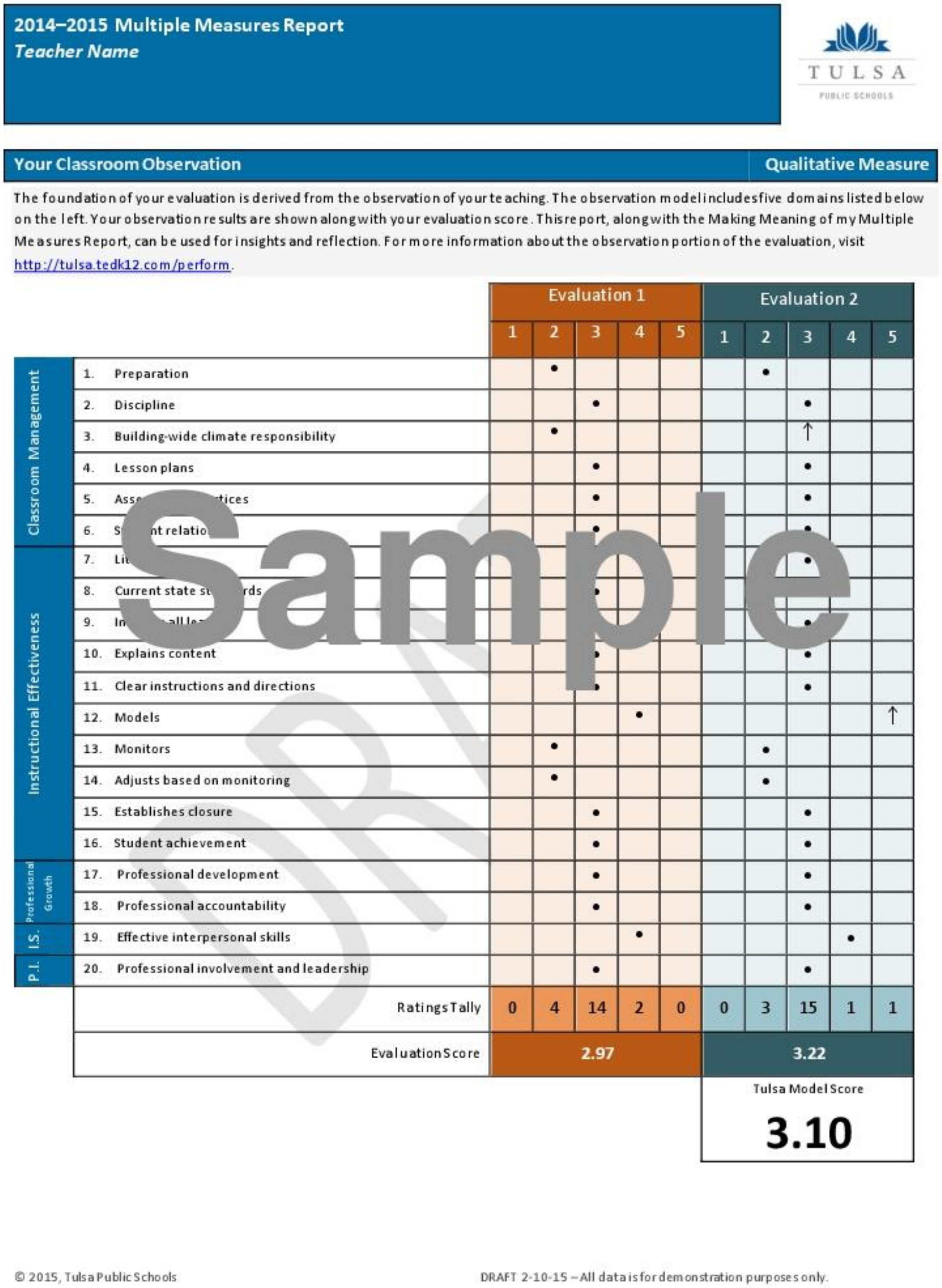
CONTACT US:
Office of Teacher and
Leader Effectiveness
ESC – 2nd Floor, Room 228
(918) 746-6223

A MESSAGE FROM DR. BALLARD

All of us at Tulsa Public Schools and the Board of Education thank you for your hard work over the past year and throughout your service. We hope you find value in our efforts to provide feedback for growth and evaluation.

If you would like to share feedback on how we can serve you better to grow our students, contact us at TheTulsaModel@tulsaschools.org

Sample



2014–2015 Multiple Measures Report
Teacher Name

TULSA
PUBLIC SCHOOLS

Your Classroom Observation | **Qualitative Measure**

The foundation of your evaluation is derived from the observation of your teaching. The observation model includes five domains listed below on the left. Your observation results are shown along with your evaluation score. This report, along with the Making Meaning of my Multiple Measures Report, can be used for insights and reflection. For more information about the observation portion of the evaluation, visit http://tulsa.tedk12.com/perform.

		Evaluation 1					Evaluation 2				
		1	2	3	4	5	1	2	3	4	5
Classroom Management	1. Preparation		●					●			
	2. Discipline			●					●		
	3. Building-wide climate responsibility		●						↑		
	4. Lesson plans			●					●		
	5. Asse[illegible]tices			●					●		
	6. S[illegible]nt relatio[illegible]			●					●		
Instructional Effectiveness	7. Li[illegible]			[illegible]					●		
	8. Current state st[illegible]rds			●					●		
	9. In[illegible]all le[illegible]			[illegible]					●		
	10. Explains content			●					●		
	11. Clear instructions and directions			●					●		
	12. Models				●						↑
	13. Monitors		●					●			
	14. Adjusts based on monitoring		●					●			
	15. Establishes closure			●					●		
	16. Student achievement			●					●		
Professional Growth	17. Professional development			●					●		
	18. Professional accountability			●					●		
I.S.	19. Effective interpersonal skills				●					●	
P.I.	20. Professional involvement and leadership			●					●		
	Ratings Tally	0	4	14	2	0	0	3	15	1	1
	Evaluation Score	2.97					3.22				

Tulsa Model Score

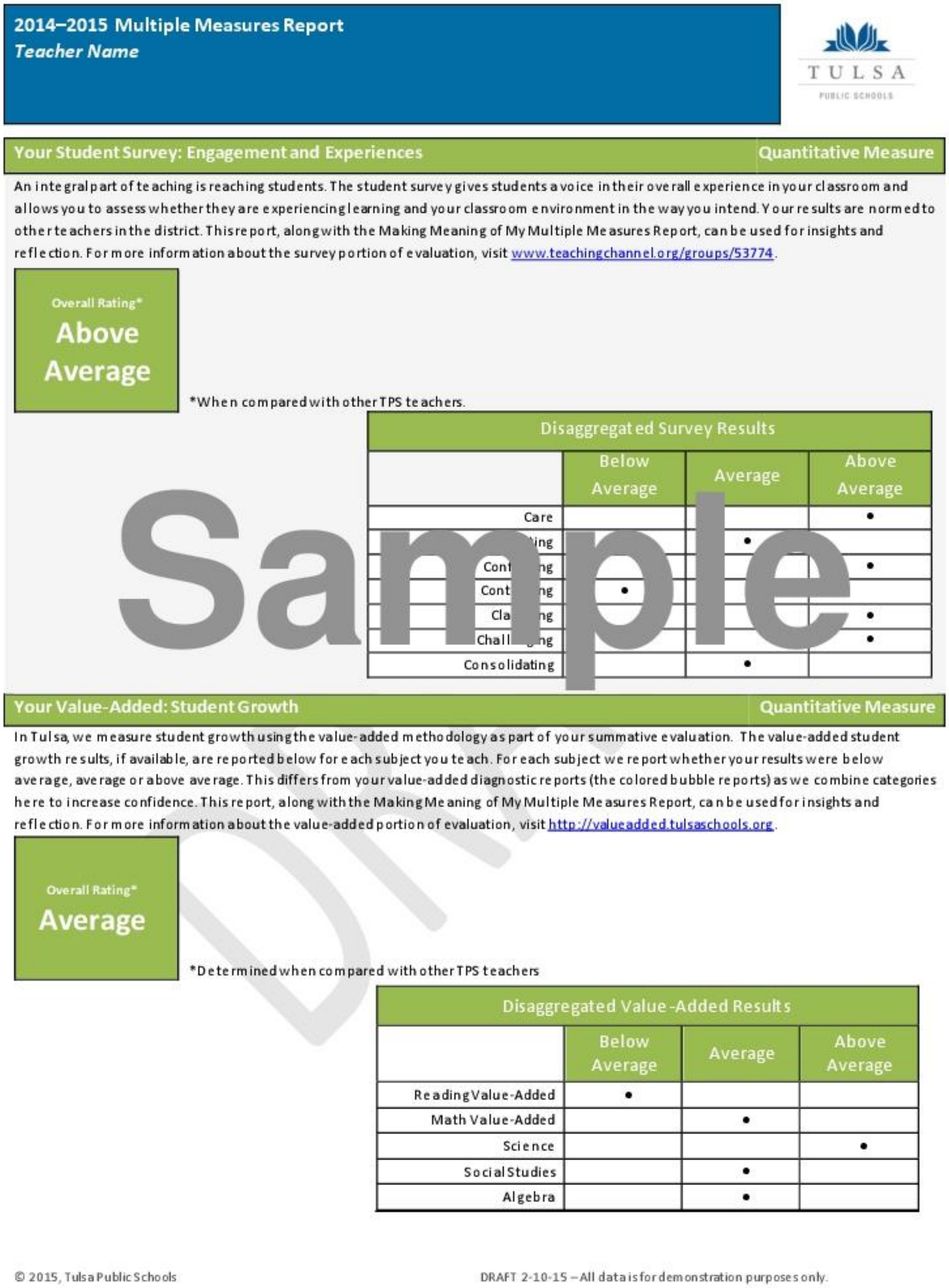
3.10

Sample

DRAFT

© 2015, Tulsa Public Schools

DRAFT 2-10-15 – All data is for demonstration purposes only.



2014–2015 Multiple Measures Report
Teacher Name

TULSA
PUBLIC SCHOOLS

Your Student Survey: Engagement and Experiences — **Quantitative Measure**

An integral part of teaching is reaching students. The student survey gives students a voice in their overall experience in your classroom and allows you to assess whether they are experiencing learning and your classroom environment in the way you intend. Your results are normed to other teachers in the district. This report, along with the Making Meaning of My Multiple Measures Report, can be used for insights and reflection. For more information about the survey portion of evaluation, visit www.teachingchannel.org/groups/53774.

Overall Rating*
Above Average

*When compared with other TPS teachers.

Sample

Disaggregated Survey Results			
	Below Average	Average	Above Average
Care			•
[illegible]ing		•	
Con[illegible]ng			•
Cont[illegible]ng	•		
Cla[illegible]ng			•
Chall[illegible]ng			•
Consolidating		•	

Your Value-Added: Student Growth — **Quantitative Measure**

In Tulsa, we measure student growth using the value-added methodology as part of your summative evaluation. The value-added student growth results, if available, are reported below for each subject you teach. For each subject we report whether your results were below average, average or above average. This differs from your value-added diagnostic reports (the colored bubble reports) as we combine categories here to increase confidence. This report, along with the Making Meaning of My Multiple Measures Report, can be used for insights and reflection. For more information about the value-added portion of evaluation, visit http://valueadded.tulsaschools.org.

Overall Rating*
Average

*Determined when compared with other TPS teachers

DRAFT

Disaggregated Value-Added Results			
	Below Average	Average	Above Average
Reading Value-Added	•		
Math Value-Added		•	
Science			•
Social Studies		•	
Algebra		•	

© 2015, Tulsa Public Schools

DRAFT 2-10-15 – All data is for demonstration purposes only.