STRATEGIC DATA PROJECT

SDP FELLOWSHIP CAPSTONE REPORT

Teacher Evaluation that Informs Professional Learning: Unpacking Classroom Observations and Student Surveys for What Matters Most for Student Achievement

Marco A. Muñoz, Jefferson County Public Schools

Dena H. Dossett, Jefferson County 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

With the continued pressure to increase student academic achievement, improving teacher effectiveness has become a national imperative. Using well-established measures of effective teaching, the focus of this study was to identify key predictors of teacher effectiveness operationalized as student growth in reading and mathematics. To accomplish the goal of improving teaching by designing high-leverage professional learning, our study will contribute by identifying the strongest predictors among all variables included in the student survey and the classroom observations.

Currently, there is ample research evidence that indicates teachers matter for student learning (Hanushek & Lindseth, 2009; Rivkin, Hanushek, & Kain, 2005; Stronge, 2007). Yet, investigating *how* teacher's impact student learning remains as a critical question. Our study is an exploration on the topic of predictors of teacher effectiveness using data associated with the new Kentucky teacher evaluation system. The state of Kentucky implemented a teacher evaluation pilot in the school year 2013–14 that informed the full implementation of the model in the school year 2014–15. The new teacher evaluation system in Kentucky is titled the Professional Growth and Effectiveness System (PGES) and has shifted away from a checklist-based model with little differentiation in ratings of teacher effectiveness (Weisberg, Sexton, Mulhern, & Keeling, 2009) to a student achievement-centered model that incorporates multiple sources of evidence to inform professional practice (Cantrell & Kane, 2013; Kane, McCaffrey, Miller, & Staiger, 2013). The new teacher effectiveness system, based on Danielson's Framework for Teaching (Danielson, 2009), Ferguson's Tripod Survey (Ferguson, 2012; Ferguson & Danielson, 2014) and Student Growth Percentiles (Betebenner, 2008), is designed to allow for more targeted and meaningful professional learning for educators in Kentucky. Nevertheless, there are still areas of exploration surrounding the new teacher evaluation system, including (a) the magnitude of correlations among the sources of evidence used for identifying teacher effectiveness and (b) the identification of high-leverage, power domains from classroom observations and constructs from student surveys that predict student growth in reading and mathematics.

Literature Review

As federal funding has become increasingly dependent on demonstrating robust systems of educator effectiveness in recent years, there has been an increase in the number of research studies on this topic. The seminal work, Measures of Effective Teaching (MET) project (Kane & Cantrell, 2010), serves as an instrumental cornerstone for establishing a common conversation around the complex topic of teacher effectiveness, and continues to shape and inform research on the factors that contribute to teacher effectiveness. The purpose of this literature review is to highlight the findings from the MET studies as well as other relevant studies associated with the key constructs used in our study, namely student perception surveys and teacher classroom observations.

Studies of Student Perceptions.

The initial study from the MET project, *Learning about Teaching*, explored the relationship between student perceptions, as measured by the Tripod survey, and teacher effectiveness (Kane & Cantrell, 2010). This ground-breaking work impacts our study, because the state of Kentucky has multiple sources of effective teaching including a student perception survey (i.e., an adaptation of the Tripod survey). The main finding was that students were able to identify teacher effectiveness based on seven constructs (7C) that are core to a student's experience in the classroom. The researchers found that student perceptions in one class predict large differences in student achievement gains in other classes taught by the same teacher, especially in the subject area of mathematics. Also, researchers found that some constructs had a stronger relationship to a teacher's value-added measures than others, including: (a) teacher's ability to manage a classroom and (b) teacher's ability to challenge students with rigorous work and effort. The student survey also helps provide evidence of effective teaching in non-tested grades and subjects where there are no value-added scores. Furthermore, the student survey is helpful because it can provide specific and actionable feedback for teacher professional learning.

A second study, *Asking Students about Teaching* (MET Project, 2012), confirmed the initial findings that students can be an important source of information on the quality of teaching and the classroom learning environment. Students can identify an effective classroom

when they experience one. A key finding of this study was that student perception surveys produce more reliable results than classroom observations or achievement gain measures. This is intuitive since student surveys aggregate the impressions of many students who have spent countless hours with a teacher. This study also emphasized the value of the survey as a feedback tool. The key challenge is to ensure that the survey instrument will produce meaningful information on teaching rather than act as a popularity contest on instructional aspects that do not matter. Equally important is to make sure that survey implementation is adequately designed, including student confidentiality, sampling, and accuracy of reporting. All of these issues can be addressed with proper piloting, clear protocols, trained coordinators, and quality control checks.

Studies of Classroom Observations.

Other studies have focused on classroom observation as a different measure of effective teaching. The studies described below examined the Danielson Framework for Teaching classroom observation tool, which is used in Kentucky's Professional Growth and Effectiveness System. Sartain, Stoelinga, and Brown (2011) conducted a two-year study of Chicago's teacher evaluation model based on the Danielson's Framework for Teaching. One of the purposes of the study was to determine if the classroom observation ratings based on the Danielson Framework were valid and reliable measures of teaching practice. The researchers concluded that three elements need to be in place to promote a successful implementation of an evidence-based teacher evaluation system: (a) a rigorous instructional rubric like the Danielson Framework to provide a common language and performance expectations, (b) a well-developed system that helps principals manage their evaluation responsibilities, and (c) professional learning opportunities for principals to develop their instructional coaching skills and engage in meaningful conversations about improvement of instructional practice.

In *Gathering Feedback for Teaching* (Kane & Staiger, 2012), MET researchers looked into examining the predictive value of classroom observations for teacher evaluation on student achievement gains. Researchers found that an accurate observation rating requires two or more lessons, each scored by a different certified observer. Achieving high levels of reliability of classroom observations requires several quality assurances: observer training and certification,

system-level audits using a second set of impartial observers, and use of multiple observations whenever stakes are high. In addition, evaluation systems should include multiple measures, not just observations or value-added measures alone. In fact, the researchers argue that combining multiple measures offers three advantages: (a) greater predictive power (slightly better than student achievement gains alone, but significantly better than observations alone), (b) greater reliability (especially when student feedback or multiple observation scores are included), and (c) the potential for diagnostic insight that allows teachers to improve their practice (which cannot be provided by student achievement gains alone). From a practitioner perspective, classroom observations and student surveys have the potential to more accurately identify strengths and address specific weaknesses in teachers' practice through intentional professional learning.

Studies with Multiple Measures Combined

The purpose of the final study of the MET project, *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study* (Cantrell & Kane, 2013), was to examine efforts to build new feedback and evaluation systems that include multiple measures designed to support teacher growth. Researchers found that student perception surveys and classroom observations can provide meaningful feedback to teachers. Furthermore, they also can help system leaders prioritize their investments in professional development to target the major gaps between teachers' actual practice and the expectations for effective teaching. Each measure adds something of value: classroom observations provide rich feedback on practice, student perception surveys provide a reliable indicator of the learning environment and give voice to the intended beneficiaries of instruction, and student learning gains can help identify groups of teachers who, by virtue of their instruction, are helping students learn more.

More recently, Chaplin, Gill, Thompkins, and Miller (2014) explored three measures of teacher effectiveness in the Pittsburgh Public Schools, namely professional practice observations, student surveys, and student growth. This study provides insight into how these factors interrelate to differentiate teacher effectiveness, and also serves as an important resource for districts and states to refine their own measures within a teacher evaluation

system. In particular, the measures included in this study were: the Research-based Inclusive System of Evaluation (RISE), based on Danielson's Framework for Teaching, 7C student survey developed by Ferguson as part of the Tripod Project, and a value-added measure of change using student test scores. Researchers found that all three measures are positively and moderately correlated, indicating that they are complementary measures of teacher effectiveness. The researchers underscore the importance of using these multiple measures beyond simply defining teacher performance, but as valuable indicators for identifying professional development needs.

In summary, past research indicates that effective teaching does make a difference in student learning and that teaching is too complex for any single measure of performance to capture it accurately. This is why identifying effective teachers requires multiple measures. The challenge is to learn from the multiple measures in ways that support professional growth and effective teaching while avoiding a narrow focus on a few aspects of effective teaching, particularly those associated with value-added measures only.

Case Study

Agency Profile

The Jefferson County Public Schools (JCPS, Louisville, KY) has two Strategic Data Project (SDP) Fellows located in the Data Management Division with both working on the Professional Growth and Effectiveness System (PGES). The SDP Fellows serve on a district-level PGES implementation team and participate in a state-level partnership with Education Delivery Institute (EDI). The district-level PGES implementation team is composed of representatives from all six achievement areas, the SDP Fellows, and Human Resource leadership. The focus of the team's work is to support and track the progress of implementing the PGES system across the entire state.

JCPS is a large urban district with 155 school sites and 6,200 teachers who serve a diverse population of 100,673 students, with 47% White, 37% Black and 7% Latino and 64% receiving free/reduced-price lunch. JCPS is proud of our student diversity and continues to be a nationally recognized integrated system. We serve students speaking over 107 different

languages and we have a system of alternative schools that ensures "there is a place for every child."

The vision of the district is that all JCPS students graduate prepared to reach their full potential and contribute to our society throughout life. The mission of the district is to provide relevant, comprehensive, quality instruction in order to educate, prepare, and inspire our students to learn. The priorities of JCPS, as expressed in the four focus areas of Strategic Plan are: (1) increased learning (goal: every student progresses in his or her individual learning), (2) graduation and beyond (goal: every student graduates prepared with enduring twenty-first century skills and dispositions for his or her postsecondary choices and life), (3) stakeholder involvement/engagement (goal: increase partnerships with parents, community, and educational organizations to enrich student learning and experiences), and (4) safe, resourced, supported, and equipped schools (goal: every educator will provide effective instruction and response to student needs). This work relates to the fourth focus area on the specific target related to effective teachers; for example, strategy 4.1.3 is focused on educator growth and effectiveness.

Research Question

Our research focused on predictors of teacher effectiveness operationalized as student growth. In particular, the study focused on the magnitude of correlations among the sources of evidence used for identifying teacher effectiveness and the identification of high-leverage, power domains from classroom observations and constructs from student surveys that predict student growth in reading and mathematics.

After conducting a comprehensive review of the literature, it seems clear that many of the future lessons regarding teacher feedback and evaluation must necessarily come from states and districts implementing teacher effectiveness systems. There are common challenges that states and districts across the nation face in developing teacher evaluation systems of both technical and adaptive nature (Heifetz, 2002). Practitioners must consider *which* measures to include, the *extent* to which each factor should be considered, and the factors of the system that impact the *reliability* of the measures. The importance of using findings from both research and practitioner communities will become increasingly critical as high stakes

consequences are associated with full implementation of teacher evaluation models. Future lessons need to be collected as states and districts innovate, assess the results, and make needed adjustments.

As a large urban school district piloting a state-level multiple measures teacher evaluation system, the goal is to build a system that will better support effective teaching in every classroom. The purpose of the current study is to advance our understanding of the relationships among the multiple measures used in the statewide professional growth and effectiveness system. The findings from the study will help provide useful information for the state and districts in refining the system as it moves towards full implementation with accountability consequences as well as better informing district decisions when designing impactful professional development. To accomplish the goal of improving teaching by designing high-leverage professional learning, the study will contribute by identifying the strongest predictors among all variables included in the student survey and the classroom observations. As SDP Fellows, there was clarity about the role of data analytics and research as related to the three SDP core domains of (a) policy and research, (b) measurement and analysis, and (c) leadership and management. For the domain of policy and research, the SDP Fellows studied the important prior work conducted by MET researchers in order to inform the research agenda and design. For the domain of measurement and analysis, working closely with the Kentucky Department of Education helped to operationalize the main variables associated with the analysis and to ensure the sample would be representative of the entire state. The SDP Fellows served an important leadership role within the district implementation team by bringing local and external research to inform critical domains associated with successful delivery, such as professional development, human resources, and technology.

Project Scope

Since this work was identified as a focus area of the district's strategic plan, it required high levels of stakeholder engagement throughout the district. The findings from this study have far-reaching implications for decision making in several divisions in the district, including professional development, human resources, management information systems, computer education support, finance, and communications. The plan was executed using a timeline that

included both short- and long-term goals. The short-terms goals of the SDP Fellows were associated with the flow of data analysis: define, clean, merge, and analyze, as well as with providing support to the weekly implementation team meetings. The long-term goals were to connect the findings with local decisions as well as contribute to national research on teacher effectiveness.

The purpose of the present study was to investigate classroom teacher observation (i.e., Danielson's Kentucky Framework for Teaching Domains) and classroom student survey (i.e., Kentucky's adaptation of Ferguson's 7C) variables that influence classroom value-added measures. The present study was designed to contribute to existing theoretical and practical knowledge about factors that influence students' value-added measures in the particular context of Kentucky. This investigation examined variables that can assist administrators and instructional coaches in targeting high-leverage professional development for improving teaching practices, which in turn, can positively impact student learning.

A correlational design was used for this investigation (Campbell & Stanley, 1963; Cook & Campbell, 1979). A correlational design is a prediction study where variables are not manipulated as in experimental research. Although correlation does not establish causality, the results of correlational studies support subsequent experimental research that allows for causal explanations. As expressed by Campbell and Stanley (1963, p. 64): "In this sense, the relatively inexpensive correlational design can provide a preliminary survey of hypotheses, and those which survive this can then be checked through more expensive experimental manipulation."

Given the predictor and outcome variables, the statistical procedure used to analyze the data was the Ordinary Least Squares (OLS) hierarchical multiple regression (Stevens, 1999). The multiple regression model and the statistical procedures used in this study were based on the recommendations by Stevens (1999) and included the following steps: (1) calculation of a correlation matrix for all the variables included in the analysis, (2) hierarchical or block entry of the sets of predictor variables into the regression equation, (3) computation of raw score regression coefficients, (4) computation of standardized regression coefficients, (5) calculation of increment in R-squared for each block entry, (6) calculation of R-squared, and (7) calculation of R-squared Adjusted.

Classroom teachers teaching English language arts and mathematics across the state of Kentucky public schools served as participants of this study. The participants, mathematic teachers (N = 194) and reading teachers (*N*=191), were considered an adequate sample based on a priori statistical power perspective. Stevens (1999) indicated that about 15 participants per predictor are needed for a reliable regression equation in the social sciences. A reliable regression equation means that the equation will cross-validate well. The participant/predictor ratio in this study was given by the following computation: 15 participants multiplied by 11 predictors result in 165 participants as a minimum sample size. The descriptive data for the study participants appear in the Appendix 1.

Predictor Variables: The present investigation studied the effects of two sets of predictor variables: student survey constructs and classroom observation domains. The first set of predictor variables were associated with a classroom-level perception survey that measures teaching effectiveness from the perspective of students. The STUDENT survey includes the seven constructs which represent a modified 7C survey (see table below for constructs and sample items). The surveys consist of 22 items with a 7-point Likert-scale at the elementary level and a 25 items on a 5 point scale at the middle school level. The results are aggregated into percent agreement with the items. The survey generates information about how students experience teaching practices and learning conditions in the classroom, as well as information about how students assess their own engagement. The descriptive information for the STUDENT survey appears in Appendix 2.

The second set of predictor variables was associated with the Kentucky-adapted Danielson's Framework for Teaching. The Kentucky Framework takes into account the Kentucky Teacher Standards, the Kentucky Board of Education's Program of Studies, Kentucky Core Academic Standards, and the Kentucky Department of Education's Characteristics of Highly Effective Teaching and Learning. The Kentucky Framework for Teaching is divided into multiple standards clustered into four domains of teaching responsibility: (1) Planning and Preparation, (2) Classroom Environment, (3) Instruction, and (4) Professional Responsibilities. Each domain is comprised of several components and teacher performance is rated for each

component according to four performance levels: Ineffective, Developing, Accomplished, and Exemplary.

Outcome Variables: The present study used classroom value-added measures, operationalized as student-growth percentiles, as the outcome variable. We are using both reading and mathematics, three-year median growth percentiles for teachers. Median Student Growth Percentile (MSGP) are calculated for classroom teachers and contributing professionals in grades 4–8 in reading and/or mathematics. The MSGP are based on the Student Growth Percentiles (SGP). SGP measures change in an individual student's performance over time. Each student's rate of change is compared to other students with a similar test score history (i.e., academic peers). SGP focuses on the relative standing of a student from year to year compared to the student's academic peers. The academic peers are students who perform very similarly on the test to the student. The student is only compared to students who start at the same place. The rate of change is expressed as a percentile. Students who outpaced their peer group would be in the percentile ranks of 50–99 while students who underperformed their peer group would be in the percentile ranks of 1–49. In Kentucky, the acceptable rank for growth is the 40th percentile. Students who score at the 40th percentile or higher are considered to have typical or higher annual growth in Kentucky. A requirement for the SGP computation is having two test scores from two different years for each student in the same subject area; in Kentucky, only reading and mathematics are tested each year from grades 3–8.

Results

The purpose of the present study was to investigate classroom teacher observation and student survey variables that influence student value-added measures in reading and mathematics. The present study addressed how student survey constructs and teacher observation domains predict value-added measurement in reading and mathematics. Appendices 3 and 4 present the correlation matrix for the reading and mathematics study respectively. The variables that correlated highest with the reading outcome variable were all associated with the classroom observation domains: planning and preparation ($r = .23$, $p = .001$), instruction ($r = .22$, $p = .001$), classroom environment ($r = .19$, $p = .004$), and professional responsibility ($r = .19$, $p = .005$). The variables that correlated highest with the mathematics

outcome variable were associated with both classroom observation domains as well as student survey constructs: classroom environment ($r = .32$, $p < .001$), instruction ($r = .30$, $p < .001$), planning and preparation ($r = .25$, $p < .001$), support ($r = .26$, $p < .001$), and discipline ($r = .26$, $p < .001$).

Hierarchical multiple regression were calculated to predict teacher effectiveness in reading and mathematics based on student survey constructs and classroom observation domains. A significant regression model was found in reading [$F (11, 179) = 5.34$, $p < .001$] and mathematics [$F (11, 182) = 5.32$, $p < .001$]. As expected, only some of the variables were significant predictors for reading and for mathematics. For reading, the majority (four out of seven) of the student survey's constructs were significant predictors: engage, nurture, transparency, and trust; however, none of the classroom observation domains were significant. For mathematics, two student survey's constructs were significant predictors: support and discipline; in addition, the domain of classroom environment was a significant predictor. Overall explained variance was approximately 20% for both reading and mathematics. Nye, Konstantopoulos, and Hedges's (2004) study on teacher effects suggested that from 7% to 21% of the variance in achievement gains is associated with variation in teacher effectiveness.

Results of the hierarchical multiple regression are displayed in Table 1. The present investigation studied the effects of two sets of predictor variables: (a) student survey constructs and (b) classroom observation domains. Table 1 shows the values for raw score regression coefficients (b), standardized regression coefficients (B), R-square, and change in R-squared values. The standardized regression coefficients represented the relative contribution of each of the predictor variables in the regression equation in terms of explaining variance on the outcome variable. The R-square value tends to be an inflated estimate of how well the regression model fits the population. The Adjusted R-Square statistic corrects the R-square to more closely reflect the goodness of fit to the population (Cohen & Cohen, 1983). The closeness in magnitude between the R-square and the Adjusted R-square indicated that the variance attributable to sampling error was small.

A key success of this work was identifying the predictive role of the student voice constructs. For example, two aspects of the student voice constructs were notably important—

discipline (control) and academic press (support). These findings were contrary to the stakeholder's expectations about the role of caring (nurturing) and captivating (engagement); however, it is important to note that these factors are still valuable for all teachers, but when it comes to student growth they have limited predictive power.

Due to the data limitations, we were not able to unpack the relative contribution of the specific components within the domains of the classroom observations. For example, within the classroom environment, it would be helpful to know which of the five components explain the most variance in student growth. This information in turn would help us target professional development opportunities for teachers. Another key challenge was working with data that had a limited distribution of teacher effectiveness. If research indicates the importance of certain factors in predicting student growth, policy makers need to recognize this work when designing teacher effectiveness classification levels and associated learning support systems.

Table 1. Hierarchical multiple regression for reading (N = 191) and mathematics (N = 194)

Variable	*B*	*β*	R^2	ΔR^2
Reading				
Step1			.18***	.14***
1. Support	-.08	-.18		
2. Transparency	.11	.30*		
3. Understand	.08	.23*		
4. Discipline	.03	.11		
5. Engage	-.15	-.55***		
6. Nurture	-.16	-.46**		
7. Trust	.16	.42**		
Step2			.25**	.20**
1. Planning and Preparation	1.17	.11		
2. Classroom Environment	-.11	-.01		
3. Instruction	1.37	.14		
4. Professional Responsibilities	1.19	.10		
Mathematics				
Step1			.15***	.12***
1. Support	.27	.31**		
2. Transparency	.10	.13		
3. Understand	-.12	-.19		
4. Discipline	.08	.17*		
5. Engage	-.07	-.13		
6. Nurture	-.09	-.15		
7. Trust	.02	.02		
Step2			.24***	.20***
1. Planning and Preparation	2.46	.13		
2. Classroom Environment	3.72	.20*		
3. Instruction	2.31	.13		
4. Professional Responsibilities	-2.20	-.11		

* p < .05; **p < .01; ***p < .001.
Reading: F (11, 179) = 5.34, p < .001
Mathematics: F (11, 182) = 5.32, p < .001

Lessons Learned

The present study was designed to contribute to existing theoretical and practical knowledge about factors that influence students' value-added measures in the particular context of Kentucky. Its purpose was to investigate classroom teacher observation (i.e., Danielson's Kentucky Framework for Teaching Domains) and classroom student survey (i.e., Kentucky's adaptation of Ferguson's 7C) variables that influence classroom value-added measures. This study used state-level *pilot-year* data from the implementation of the new teacher evaluation system without accountability consequences. The findings from the study help provide useful information for the state and districts in refining the system as it moves towards full implementation with accountability consequences as well as better informing district decisions when designing impactful professional development. To accomplish the goal of improving teaching by designing high-leverage professional learning, our study contributes to this discussion by identifying the strongest predictors among all variables included in the student survey and the classroom observations.

This predictability study, focusing on student achievement as outcome variable, helps us identify the most important factors from the Danielson Framework for Teaching and the Ferguson STUDENT voice survey that contribute to improving student learning measured using the Median Student Growth Percentiles for three combined years of test scores.

In reading, all Danielson's Framework for Teaching domains were significantly correlated with student achievement. Only engagement was significantly correlated with student achievement, but in a negative way; perhaps the more effective teachers do not necessarily make learning more interesting and just focus on delivering their lesson with stoicism—doing what they need to do to ensure student learning, regardless of being interesting or not.

In mathematics, all Danielson's Framework for Teaching domains were significantly correlated with student achievement. Support, discipline, and trust were significantly correlated with student achievement in a positive way. The first two are considered press factors since support and discipline are about rigor and effort in the context of managing instructional time in an efficient manner; the third factor, trust, is also very important since the

press factors can be successful in a trusting environment or where positive relationships exist between teachers and students.

This study regression findings for mathematics achievement basically corroborated prior research. The MET study, *Learning about Teaching* (2010), researchers found that some constructs had a stronger relationship to a teacher's value-added measures than others: teacher's ability to control a classroom and teacher's ability to challenge students with rigorous work. Classroom observations of the environment (e.g., discipline) was the significant predictor and this provided additional support to the value of discipline (as expressed by students in the survey).

The findings in reading could not be linked to prior research and were puzzling to us. On one hand, the sign of two regression coefficients were negative: engagement and nurture; our hypothesis is that classrooms with high engagement and nurturing not necessarily provide the necessary discipline for keeping the academic press (i.e., time on task) that is found in high-achieving classrooms. On the other hand, trust and transparency were regression coefficients with positive signs; our hypothesis is that teachers who build trust (i.e., confer) and are able to have enough transparency (i.e., clarity) help students understand the complexities of reading.

Great teaching makes a difference. Teaching is too complex for any single measure of performance to capture it accurately. Identifying great teachers requires multiple measures. Perhaps states and districts need to embrace multiple measures for targeted feedback and to support decision-making. The challenge is to look at multiple measures in ways that support effective teaching while avoiding such unintended consequences as too-narrow a focus on one aspect of effective teaching.

Implications for Professional Learning and Policy

Planning and preparation before delivering instruction, having a classroom management system in place, delivering the lesson, and exercising professionalism matter The Danielson's Framework for Teaching sequence was well correlated with student achievement for both reading and mathematics. We need to continue supporting teacher preparation programs and in-service professional development that helps master the four essential domains of effective teaching.

In terms of the STUDENT voice survey, it is recommended that professional learning focus on the twin press constructs of support and discipline. Pressing for effort and rigor in the context of an efficient use of instructional time (i.e., avoiding disruptive behavior) can make a significant contribution to student learning. Since student surveys shows more differentiation than adult observations, policy makers need to consider if student voice should be included in the state's teacher effectiveness system. Currently, these survey constructs are only considered evidence for teachers' professional practice, but it is not a separate dimension of the teacher evaluation system.

Perhaps well-rounded educators need to strive to become knowledgeable on both theoretical frameworks. More importantly, applying the domains and constructs is the instructional challenge. It is very likely, for example, that some survey constructs might vary in their level of implementation depending on the particular needs of different cohorts of students.

Study Limitations and Implications for Future Research

As it is the case for all research, the results of this study must be interpreted within certain limitations. The present study is limited due to sampling procedure (i.e., pilot implementation year with no accountability), nature of the correlational design (i.e., correlation does not equal causation), the lack of distribution on teacher effectiveness (i.e., sample had no *ineffective* teachers and only approximately 5% *developing* teachers), and focusing in reading and mathematics in grades 3–8 (i.e., other subjects and grades might respond differently to the surveys and observations). In general, this study is limited by the problems of using student test scores to evaluate teachers (Baker, Barton, Darling-Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, & Shepard, 2010).

Given the limited sample of teachers with student growth percentiles, future research needs to explore whether classroom observation and student voice variables are equally important when predicting student learning objectives which applies to the majority of teachers. Additionally, future research needs to study the individual components associated with the Danielson Framework for Teaching to go beyond the macro-level analysis utilized by this study (i.e., domains). Future research might also analyze the data with clustering

techniques, such as Hierarchical Linear Modeling (HLM), to distinguish between individual teacher level from school and district level variance. Additional covariates related to teachers (e.g., demographics, education, certification) and students (e.g., poverty, disability, language proficiency) might also help understand teacher effectiveness in a more interactive way.

Conclusion

In a way, this study conducted in the state of Kentucky is just the beginning of a research agenda associated with professional growth and teacher effectiveness. More firm conclusions can be derived from future studies associated with full implementation of the new model, particularly once accountability consequences will be established across the commonwealth. For now, we can confidently reaffirm our desire to place an effective teacher in every classroom (Darling-Hammond, 1996; Kane, T. J., Kerr, K. A., & Pianta, R.C., 2015; Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L., 2010). An effective teacher that has the disposition, knowledge, and skills associated with the four domains of the Danielson Framework for Teaching, including (a) planning instruction, (b) managing the classroom environment, (c) delivering teaching while assessing, and (d) exercising professionalism inside and outside of the classroom. An effective teacher that demonstrate to students the disposition, knowledge, and skills associated with the Ferguson STUDENT voice survey constructs, particularly when it comes to pressing for effort and rigor in the context of well-managed instructional use of time. Although we know not even multiple measurements are likely to capture the complexity of measuring teaching, we need to continue our quest to identify effective classroom practices that will allow our students to be college and career ready when finishing their K–12 experience.

References

Baker, E. L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers.* Washington, D.C.: Economic Policy Institute.

Betebenner, D. W. (2008). A primer on student growth percentiles. *Dover, NH: National Center for the Improvement of Educational Assessment.* Retrieved February, 18, 2011.

Campbell, D. T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. Seattle, WA: *MET Project Research Paper.*

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools* (REL 2014–024). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin.

Danielson, C. (2009). A framework for learning to teach. *Educational Leadership*, *66*(9).

Darling-Hammond, L. (1996). What matters most: A competent teacher for every child. *Phi Delta Kappan. 78(3)*, 193–200.

Ferguson, R. F. (2012, November). Can Student Surveys Measure Teaching Quality? *Phi Delta Kappan*, 94(3), 24–28.

Ferguson, R. F. & Danielson, C. (2014). How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching. *Designing Teacher Evaluation*

Systems: New Guidance form the Measures of Effective Teaching Project by Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta (Eds.). Thousand Oaks, CA: Jossey-Bass.

Hanushek, E. A., & Lindseth, A. A. (2009). *Schoolhouses, courthouses, and statehouses: Solving the funding-achievement puzzle in America's public schools*. Princeton, NJ: Princeton University Press.

Heifetz, R. A., & Linsky, M. (2002). *Leadership on the line.* Boston, MA: Harvard Business School Press.

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. Seattle, WA: *Bill & Melinda Gates Foundation*.

Kane, T. J., & Cantrell, S. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. Research Paper. MET Project. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., Kerr, K. A., & Pianta, R.C. (2014). Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project. San Francisco, CA: Jossey-Bass.

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. Research Paper. MET Project. Seattle, WA: Bill & Melinda Gates Foundation.

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. Seattle, WA: Bill & Melinda Gates Foundation.

MET Project (2012). Asking students about teaching: Student perception surveys and their implementation. Seattle, WA: Bill & Melinda Gates Foundation.

Nye, B., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26(3)*, 237–257.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

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. Research Report. Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637.

Stevens, J. (1999). *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Stronge, J. H. (2007). *Qualities of effective teachers*. Alexandria, VA: ASCD.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review, 100(2),* 256–60.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.

Appendix 1. Participants with reading (N = 191) and mathematics (N = 194) growth scores for three years

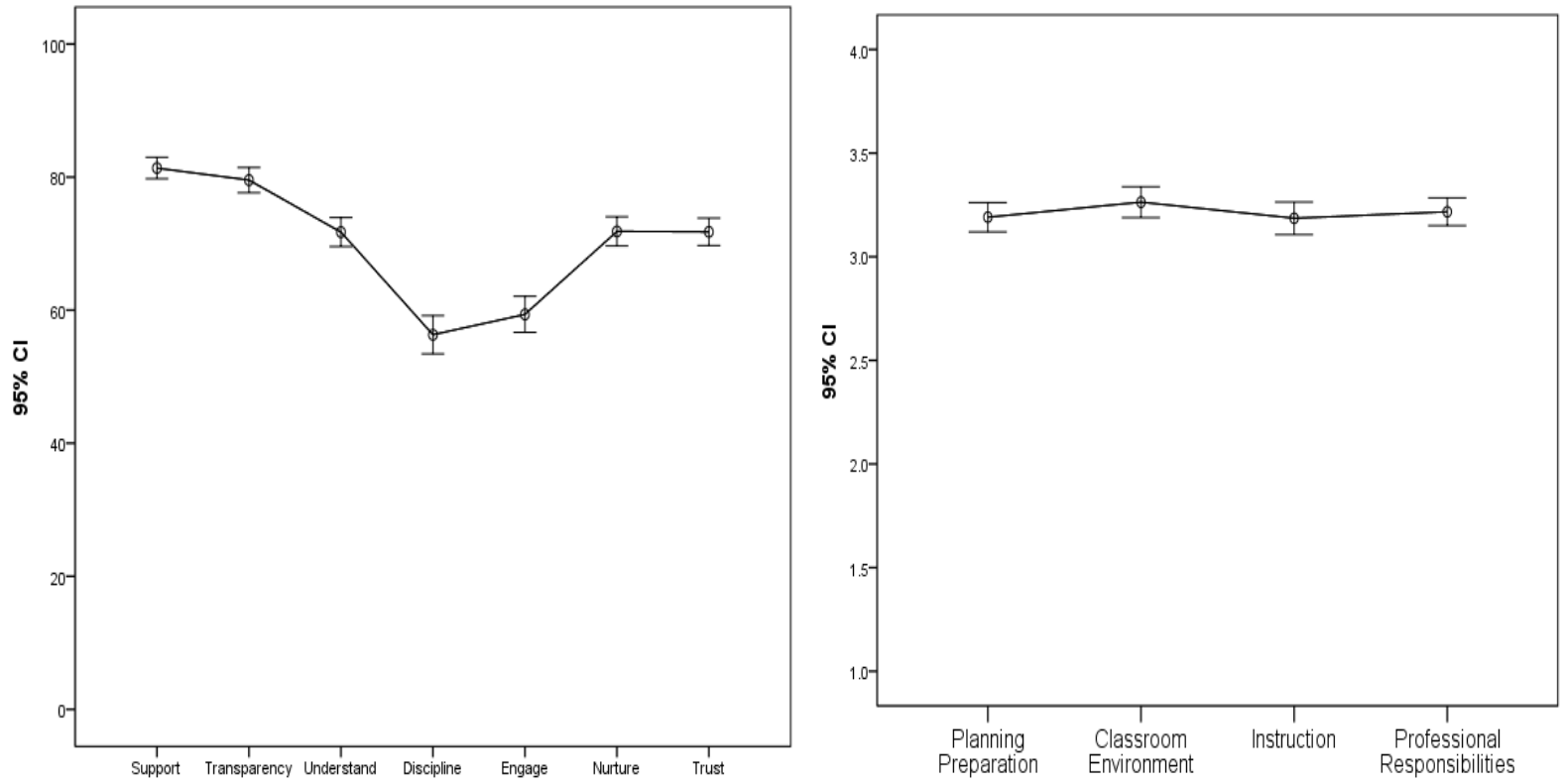
	Sample		Population	
Variable	N	Percent	N	Percent
Reading Participants				
School Level				
Elementary (Grades 3–5)	95	49.74%	329	50.62%
Middle (Grades 6–8)	96	50.26%	321	49.38%
District Type				
Rural	40	48.19%	84	52.83%
Town	27	32.53%	50	31.45%
Suburban	11	13.25%	19	11.95%
Urban	5	6.02%	6	3.77%
District Accountability				
Distinguished	19	22.89%	35	22.01%
Proficient	36	43.37%	71	44.65%
Needs Improvement	28	33.73%	53	33.33%
Overall Teacher Effectiveness Rating				
Ineffective	0	0.00%	0	0.00%
Developing	9	4.79%	111	6.18%
Accomplished	125	66.49%	1234	68.71%
Exemplary	54	28.72%	451	25.11%
Math Participants				
School Level				
Elementary (Grades 3–5)	97	50.00%	265	46.74%
Middle (Grades 6–8)	97	50.00%	302	53.26%
District Type				
Rural	43	48.86%	84	52.83%
Town	31	35.23%	50	31.45%
Suburban	11	12.50%	19	11.95%
Urban	3	3.41%	6	3.77%
District Accountability				
Distinguished	23	26.14%	35	22.01%
Proficient	35	39.77%	71	44.65%
Needs Improvement	30	34.09%	53	33.33%
Overall Teacher Effectiveness Rating				
Ineffective	0	0.00%	0	0.00%
Developing	9	4.69%	111	6.18%
Accomplished	131	68.23%	1234	68.71%
Exemplary	52	27.08%	451	25.11%

Appendix 2. STUDENT Voice Survey Constructs

Construct (STUDENT/7C)	Description	Sample Item	Reliability*
Support/Challenge	Press for effort and rigor	*My teacher wants me to explain my answers – why I think what I think*	.75
Transparency/Clarify	Explanations are clear	*My teacher explains difficult things clearly*	.78
Understand/Consolidate	Ideas get connected and integrated	*My teacher takes the time to summarize what we learn each day*	.77
Discipline/Control	Press for cooperation and peer support	*Our class stays busy and doesn't waste time*	.81
Engaging/Captivate	Learning seems interesting	*My teacher makes learning enjoyable*	.74
Nurture/Care	Encouragement and support	*My teacher in this class makes me feel that s/he really cares about me*	.79
Trust/Conferring	Students sense their ideas are respected	*My teacher wants us to share our thoughts*	.79

*Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In Designing teacher evaluation systems by Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta (Eds.). San Francisco, CA: Jossey-Bass.

Appendix 3. Error Bars with 95% Confidence Intervals



NOTE: The student voice constructs range from an average percent agreement of 56.3 (SD=20.2) for discipline to 81.4 (SD=11.3) for support. The Framework for Teaching (FFT) domains range from 3.19 to 3.26, with standard deviations around .5. Student data presents more variability. More differentiation on dimensions of teacher practice was observed from the student perspective when compared to the administrator observations.

Appendix 4. Correlations for Reading (N = 191)

	M	SD	DV	1	2	3	4	5	6	7	8	9	10	11
DV = Reading	50.73	5.45	1	-.08	-.07	-.08	.03	-.20**	-.17*	.03	.23**	.19**	.22**	.19**
1. Support	81.43	12.04		1	.70***	.64***	.52***	.60***	.68***	.72***	.14*	.09	.18**	.02
2. Transparency	78.72	14.37			1	.65***	.47***	.71***	.85***	.75***	.07	.10	.18**	.06
3. Understand	71.62	16.30				1	.56***	.79***	.73***	.71***	.11	.05	.10	.05
4. Discipline	54.68	20.52					1	.58***	.51***	.63***	.14*	.19**	.15*	.10
5. Engage	59.85	20.38						1	.77***	.76***	.10	.10	.19**	.09
6. Nurture	72.28	16.02							1	.78***	.09	.08	.17*	.04
7. Trust	72.05	14.42								1	.16*	.17*	.27***	.10
8. Planning	3.20	.51									1	.41***	.51***	.50***
9. Environment	3.31	.52										1	.61***	.53***
10. Instruction	3.20	.54											1	.48***
11. Professional	3.24	.47												1

Note: * < .05; ** < .01; *** < .001

Appendix 5. Correlations for Mathematics (N = 194)

	M	SD	DV	1	2	3	4	5	6	7	8	9	10	11
DV = Mathematics	50.94	9.73	1	.26***	.17**	.06	.26***	.06	.07	.20**	.25***	.32***	.30***	.16*
1. Support	81.36	11.30		1	.67***	.65***	.53***	.61***	.63***	.73***	.16*	.12	.10	.09
2. Transparency	79.54	13.37			1	.66***	.48***	.66***	.81***	.71***	.08	.14*	.15*	-.02
3. Understand	71.73	15.38				1	.57***	.79***	.69***	.64***	.10	.09	.11	.06
4. Discipline	56.30	20.22					1	.57***	.40***	.58***	.18**	.24***	.23**	.16*
5. Engage	59.36	19.14						1	.68***	.68***	.10	.10	.15*	.07
6. Nurture	71.84	15.45							1	.70***	.01	.11	.08	-.05
7. Trust	71.77	14.48								1	.17**	.12	.21**	.08
8. Planning	3.19	.50									1	.32***	.49***	.62***
9. Environment	3.26	.53										1	.54***	.44***
10. Instruction	3.19	.55											1	.48***
11. Professional	3.22	.47												1

Note: * < .05; ** < .01; *** < .001