MEASURING TEACHING EFFECTIVENESS USING VALUE-ADDED AND OBSERVATION RUBRIC SCORES

Andrew McKenzie, B.A., M.A.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2014

APPROVED:

Jeanne Tunks, Major Professor
Mariela Nuñez-Janes, Committee Member
David Molina, Committee Member
Prathiba Natesan, Committee Member
Jamaal Young, Committee Member
James Laney, Chair of the Department of Teacher Education and Administration
Jerry Thomas, Deanof the College of Education
Mark Wardell, Dean of the Toulouse Graduate School

McKenzie, Andrew. Measuring Teaching Effectiveness Using Value-Added and Observation Rubric Scores. Doctor of Philosophy (Curriculum and Instruction), December 2014, 163 pp., 11 tables, 6 figures, references, 96 titles.

This mixed-methods study examined the extent to which teacher performance and student performance measures correlated, and to understand which specific practices of mathematics teachers in Grades 3-5 related to student performance. Research was conducted at five elementary schools in a large, urban north Texas school district. Data sources included component scores and recorded evidence from observation rubrics, interviews with campus administrators, and value-added modeling (VAM) student growth scores.

Findings indicated a modest relationship between teacher performance levels and student performance levels. Lack of access to individual teacher VAM data, per district policy, might have impacted the strength of the relationship. Interviews with administrators and an examination of the evidence cited in the observation rubrics identified specific practices associated with highly rated mathematics teaching. Differences in administrators' experience levels with both mathematics instruction and the observation instrument might have influenced rubric scores and the level of specificity shown in evidence statements.

ACKNOWLEDGEMENTS

I would like to first thank my committee members for their thoughtful guidance throughout this process: Dr. David Molina, Dr. Prathiba Natesan, Dr. Mariela Nuñez-Janes, and Dr. Jamaal Young. I am especially indebted to my committee chair, Dr. Jeanne Tunks, for her willingness to coach at every turn, and for her incisive feedback, which vastly improved this dissertation.

I want to thank my amazing wife, Dana McKenzie, whose love, encouragement, and unswerving support were all vital to my completing this program. I'd also like to remember and recognize my parents, Jim and Dale McKenzie, who taught me the love of learning and always championed whatever I studied, from medieval philosophy to Russian literature. No doubt, they'd be just as enthusiastic about curriculum and instruction.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Measuring effective teaching has dominated recent conversations and initiative agendas in American education. The U.S. Department of Education has designed entire competitive grant programs, including the Teacher Incentive Fund (TIF) and Race to the Top (RTTT), around measuring teaching effectiveness, recognizing the most effective practitioners, and compensating them differently based on their demonstrated effectiveness (Hannaway & Mittelman, 2011). The two primary metrics of teaching effectiveness currently employed by these grant programs – multiple classroom observations by multiple observers (Danielson & McGreal, 2000) and value-added modeling (VAM) measures of student progress – were adopted as intentional departures from more traditional methods that have long been perceived as insufficient.

For decades the effectiveness of a teacher was measured by a single annual state-mandated formal classroom observation. Regardless of the specific observation instrument used, the vast majority of American teachers (over 90%) were annually rated as effective or highly effective (Weisberg, Sexton, Mulhern, & Keeling, 2009). These glowing instructional ratings, however, did not necessarily translate into positive student performance. It was common to find a school where teachers were universally rated as effective and students were universally under-achieving, or even failing. Both TIF and RTTT have begun to address the observation component of rating teacher effectiveness by implementing multiple observations by multiple observers throughout the school year and by urging districts to adopt alternative rubrics that focus on improving teaching performance, not merely evaluating it. Charlotte Danielson’s framework for teaching

(FFT) method is widely used for this purpose. It was designed as a framework of instructional components to provide a common language to educators to discuss their practice (Danielson, 2007). Measuring teacher instructional practice, however, is only one half of the effectiveness equation; measuring student performance, specifically student growth, is the other.

Student performance has typically meant one thing – achievement. Achievement, in turn, has typically meant only passing rates on annual state standardized tests in the core curricular content areas, primarily mathematics and reading (U.S. Congress, 2001). Federal and state accountability systems have focused almost exclusively on student passing rates. The TIF and RTTT initiatives have acknowledged that passing rates specifically, and student achievement generally, do not fully capture how students are performing, nor do they reflect the impact that teachers, schools, and instructional programs are having. Student progress or growth measures, provided by VAM analyses, help to complete the picture on student performance. It’s possible now to recognize that a cohort of students that, by and large, did not achieve passing standards on a state test, nevertheless showed expected progress or even greater than expected progress, according to VAM (Sanders & Horn, 1994). Conversely, it is possible to recognize when a cohort of students that, by and large, achieved or surpassed passing standards on a state test, did not show expected progress, according to VAM. In both scenarios, VAM helps to contextualize achievement data and provide a more complete student performance profile.

Statement of the Problem

Measuring teacher effectiveness through the dual channels of measuring teacher instructional performance through multiple observations conducted by multiple observers and measuring student growth performance with VAM is a departure from traditional methods that frequently led to obviously incongruent indications of effectiveness. This study has attempted to gauge whether and to what extent these methods provide consistent indications of effective teaching.

Statement of the Purpose

The purpose of this study was to investigate the extent to which teacher performance and student performance measures are correlated, and to discover which specific practices of mathematics teachers rated highly on both measures may have led to increased student performance.

Research Questions

1. What is the relationship between students' progress scores in mathematics, as measured by VAM, and their teachers' instructional performance scores, as measured by the FFT?
2. What is the relationship between levels of teacher and student performance, as articulated in interviews with campus administrators, and an examination of their documented observation evidence using the FFT?

Hypotheses/Assumptions

1. It was hypothesized that there would be a significant relationship at the $p < .05$ level between teacher performance FFT scores and student performance VAM scores.

2. It was assumed that students' value-added progress measures would positively relate to teacher instructional performance ratings. Specifically, it was assumed that teachers rated high according to their FFT observation scores would also be rated high according to their VAM student progress scores, and teachers rated low on FFT would also be rated low on VAM. In cases where the ratings do not match, any number of factors might account for the mismatch, including the possibility that one of the underlying suppositions might be flawed, namely that these are, in fact, both measures of effective teaching. It was also assumed that interviews with school administrators and an examination of their observation documentation would provide insight into confirmatory or discrepant VAM/FFT relationships. Finally, it was hoped that this study might yield a set of best teaching practices by examining the most highly rated instructional practices of those teachers who scored high on both measures.

Theoretical/Philosophical Rationale of Study

Post-Positivism/Interpretivism

Central to a post-positivist epistemology, following Thomas Kuhn (1970) and Karl Popper (1959), is the acknowledgement of ontological relativism and the placement of knowledge within particular subjective perspectives. The pursuit of knowledge through scientific research is both worthy and necessary in order to learn what we can about reality, even if that knowledge is ultimately incomplete and imperfect. Interpretivism, a post-positivist paradigm within the social sciences, emphasizes understanding a phenomenon, rather than explaining it with the intent of predicting future iterations (Charmaz, 2010). Research from an interpretivist viewpoint is focused on the way people interpret and make sense of their experiences, and the researcher applies inductive logic to distill meaning from the collected data (Grbich, 2010; Teddlie & Tashakkori, 2009; Young, 2009). When one of the instruments of research is a person (as it was in this study two-fold), the post-positivist acknowledges that the observer's own experiences and values can influence the observation process and what is observed (Henderson, 2011). At the primary level in this study, administrators observed, interpreted, and gave meaning to the instructional practices of their teachers. At a secondary level, the researcher has interpreted and given meaning to the evidence and explanations of the observers. Each layer of interpretation underscores the abstractness of the idea of teacher or teaching effectiveness.

Erickson (1986) noted that teacher effectiveness, from an interpretivist perspective, "is a matter of the nature of the social organization of classroom life – what we have called the enacted curriculum – whose construction is largely, but not exclusively, the responsibility of the teacher as instructional leader" (p. 133). Erickson's point is that the teacher is not the exclusive determinant of what happens in the classroom; rather, the classroom is a sophisticated social system as dependent on the students as it is on the teacher as they continually negotiate what will transpire in the teaching and learning exchange. It was assumed at the outset of this study that effective teaching is a construct that is sourced from a number of measures and perspectives, which is why a mixed methods approach has been used. It's also not a stretch to suggest that VAM, while clearly a quantitative method, also aligns within an interpretivist paradigm since growth is always defined within local contexts. What counts as expected growth for a cohort of students is dependent on what they have already demonstrated they are capable of achieving; growth targets differ with school contexts. At best, both the VAM and FTT measures provide signals or indicate instances of what might be considered teaching effectiveness, though it is understood that they can never fully reveal the essence of effective teaching.

Definition of Terms

VAM – In this study, VAM data are produced through the application of the Educational Value-Added Assessment System (EVAAS) methodology. Mean actual performances of a given cohort of students on a standardized assessment are compared to the mean predicted performances of the cohort in order to determine the extent of growth from one grade level to the next.

Expected growth – Expected growth for a given cohort of students on VAM is a mean normal curve equivalent (NCE) gain that is close to zero and also within the range of the standard error (both above and below zero) produced by the VAM calculation.

Greater than expected growth – Greater than expected growth, as measured by VAM, is a mean NCE gain that is positive and greater than the standard error.

Less than expected growth – Less than growth, as measured by VAM, is a mean NCE gain that is negative and less than the standard error.

FFT – The teacher observation rubric developed by Charlotte Danielson, the framework for teaching (FFT), was adopted by the researched district to form the basis of its alternative observation instrument.

ID&E – The Individual Development and Evaluation scorecard (ID&E) is the generic name for the alternative teacher observation instrument that TIF grant districts were required to adopt. What specifically they chose to use was decided locally. The researched district adopted the FFT and used 13 of its 22 components as its ID&E scorecard.

Proficient – The Danielson FFT observation rubric has four levels of performance to describe teaching practice: unsatisfactory, basic, proficient, and exemplary. The proficient level, though the third out of four levels of performance, according to Charlotte Danielson, describes consistent, effective instruction.

High T/high V – A teacher rated as high T/high V was rated high in both measures. When a teacher's overall observation score average is above his/her campus mean score, it is considered a high T (teaching performance) score. When greater than expected growth, as measured by VAM, is attributed to a teacher, it is considered a high V (value-added) score.

Limitations

Internal Validity

Causation conditions. This study was not designed to examine causation. However, it may be seen as a partial investigation on the conditions of causation. Since state standardized tests are given at the end of the school year, VAM data are generated after observation rubric data, which are generated throughout the school year. If strong relationships were found between the two measures, then we might be tempted to see particular teacher instructional performances as possible determinants of particular student performances. This relationship cannot be established in this study.

Instrumentation. VAM data in the district where the study was conducted are only reported at the grade level/content area cohort level, e.g. 4th grade mathematics. Through an abundance of caution, the district has set a policy whereby teacher-level value-added data are not reported. So, all mathematics teachers at a given grade level are given the same VAM score. In some cases, there was a single teacher who taught all mathematics classes at a grade level, so the VAM cohort score was essentially a teacher-level score. In most cases, however, more than one teacher shared the same VAM score for the entire grade level cohort.

Attrition. Common in the observed schools is student transiency. It could be the case that the students who tested at the end of the year and whose scores are figured into VAM data were not identical to the students on the roster during the observations that contributed to the rubric score data.

External Validity

Population. Generalization of the study's results to wider populations is tentative due to the narrowness of its focus on only five elementary schools. The conditions that led to these schools being included in the district's TIF grant cohort, namely that they are historically hard to staff and under-performing, might also compromise the ability to generalize from this study.

Accessible population. The ten school administrators (two per campus) who conducted the observations used in this study had a variety of familiarity with the Danielson FFT rubric, which was introduced as an observation instrument when the implementation of the TIF grant began. This study was conducted during the third year of the grant implementation. Five of the administrators were in their third year of using the FFT, two were in their second year, and three were using it for the first time.

Reactivity. This was a twofold concern. First, according to local grant implementation policy, the two main observations that produce the FFT scores were scheduled in advance. While observers had the leeway to include evidence from previous unscheduled observations, the scheduled observations formed the core of the evidence source. Teachers, then, may have been inclined to perform and act in ways that were not representative of their typical instructional performances and actions. Secondly, the ten participating administrators who took part in interviews about the FFT observations they made, the evidence they noted, and the scores they gave, might also have been influenced by the awareness that they were participating in a study.

Another potential limitation was researcher bias. As the researcher, I brought both familiarity with the content area, elementary mathematics, and detailed knowledge of the TIF grant to the study. Specifically, I served as the project manager for TIF grant implementation within the researched district. Though I did not directly supervise the participating administrators, they were very familiar with my role, and this might have had an influence on reactivity during the interviews, especially. Additionally, I previously served as a content specialist and instructional coach in elementary mathematics for the district. This background knowledge likely impacted my interpretation of the findings, even though I worked hard to neutralize my bias.

Significance of Study

This study can contribute to the conversation on measuring teacher effectiveness, particularly within the ranks of upper elementary mathematics instruction. If classroom observation rubric scores of teacher performance correlate to VAM scores of student progress, and if a set of best practices that are characteristic of teachers rated high in both measures can be determined, a plausible coaching plan might be designed and developed based on those practices.

The introduction of VAM has seen a near-simultaneous use of its data in making high-stakes decisions regarding employment, placement, and compensation. The results of this study can contribute to the debate about the appropriateness of VAM's usage in such decision-making. Before campus and district leaders begin hiring, firing, staffing, and differentiating compensation based on teachers' VAM scores, it would be wise to reflect on multiple studies – this being only one of them – that attempt to glean what VAM data are revealing about teachers and teaching.

Chapter Summary

Efforts to measure effective teaching have fueled numerous high-profile federal grant programs in recent years. This study examined the relationship between measures of teacher performance and student performance within five elementary schools. With a specific focus on mathematics instruction in Grades 3-5, one of the goals of this study was to emerge with a set of best instructional practices that might be codified and shared within the field.

CHAPTER 2

RELATED LITERATURE

The Emergence of Value-Added in Education Reform

Ordered in 1989 and passed in 1990, not only did the Kentucky Education Reform Act (KERA) open a new chapter in sweeping large-scale education policy change, it also unwittingly launched the effort to measure teaching effectiveness that continues today, fueled by a myriad of federal grant programs. It's not hard to connect a fairly straight line of dots from KERA to the No Child Left Behind Act of 2001 (NCLB) and its call for highly qualified teachers in every American classroom. Two of the most persistent questions over the past decade since NCLB have been: What is effective teaching, and how do we measure it?

Teacher effectiveness was not central to the initial conception of KERA. With this act, Kentucky executed a complete overhaul of the state public education system due to its overall ineffectiveness and rampant inequities. Chief among KERA's reforms was leveling the fiscal playing field by establishing minimum per-student expenditure thresholds across the state and offsetting district shortfalls from low tax-yields with state funds. As a result, state spending on education rose by one-third from 1990 to 1992 (Weston & Sexton, 2009). A call for instruction aligned to state curricular standards led to the establishment of the Kentucky Instructional Results Information System (KIRIS), which would develop assessments for the statewide instructional standards and provide a unified approach to harvesting assessment data and analyzing them. A bold and far-reaching goal was set under KIRIS for all schools to achieve an accountability index of 100 (meaning 100% of their students would minimally reach passing standards in all subjects) by 2012, some twenty years in the future (Innes, 2010). While the target date was re-set to 2014 based on slow progress, the precedent had been set for shaking up an entire educational system and setting specific, time-bound targets with success being measured in the currency of student test scores.

Logistical innovations introduced by KERA included ungraded primary classes, a focus on the use of portfolios to track and assess student progress, and open-ended test questions. A turn away from regimented lessons on isolated skills such as phonics or arithmetic yielded a more holistic approach to teaching and learning. All of these new instructional expectations were turned over to local school leaders, who were ordered to fill their teaching vacancies with educators capable and willing to carry out the new plans. Other systems were put in place to support school communities including family resource centers and provide opportunities for parents to become more involved in the educational setting. By the early 2000s, Kentucky students were closer to national averages in all subjects on the National Assessment of Educational Progress (NAEP) (Weston & Sexton, 2009). KERA seemed to have worked.

There were differing viewpoints, however. Innes (2010) noted, “KERA in many ways was a massive, statewide experiment, conducted at public expense, using at least a generation of students as its subjects” (p. 7). The implication is that it was not an entirely successful “experiment” and much had been wagered on it. While it is beyond the scope of the present discussion to decide whether and to what degree KERA succeeded, it cannot be denied that KERA re-cast the idea of how to assess teaching. Kentucky teachers were charged with the task of crafting lessons aligned to state standards and they understood that annual assessments would reveal how well they had taught. Collectively, the entire state educational system had set inspiring, if not daunting, goals for itself, and all teachers were expected to do their part to help reach them. Progressively inflated scores on KIRIS assessments, the constriction of the curriculum toward test-preparation, and an “undesirable narrowing and corruption of instruction” (Koretz & Barron, 1998, p. 121) were accepted as necessary, if unpleasant, side effects of this pursuit.

Two short years after KERA was introduced, Tennessee, passed the Education Improvement Act of 1992 (EIA) which increased state funding for education and required the use of accountability measures based on student performance outcomes to gauge the effectiveness of districts, schools, and teachers (“TVAAS Resources,” 2012). To measure the impact of teachers and schools on student progress, Tennessee turned to a methodology created by one of its own, William Sanders, a professor at the University of Tennessee, Knoxville. Sanders had been experimenting since the mid-1980s with ways to statistically isolate the effect that teachers have on the progress of students under their charge. The teacher effect was termed “value,” and the calculations attempted to estimate the amount of value that teachers added to their students’ progress. When this methodology was chosen for statewide implementation under EIA, it became known as the Tennessee Value-Added Assessment System, or TVAAS.

The value-added methodology underlying TVAAS uses scores from standardized tests, typically given annually to students, as its raw material. While most educators have become conditioned to examine achievement data, or student attainment rates, value-added provides progress data, or student growth rates (“Battelle for Kids”).

Achievement data are contextualized by comparisons to passing standards; value-added data are contextualized by comparisons to the historical test data of the students in question. More simply, value-added modeling, or VAM, compares students only to themselves, not to passing standards that apply equally to all test-takers.

VAM is predicated on comparing actual test performances with predicted test performances, or targets, in order to measure growth. Growth targets for a cohort of students are determined by their collective testing history. For example, if the mean normal curve equivalent (NCE) of a fourth grade math class is 53, based on their test performances in third grade, then their minimal target for their fourth grade test would be a cohort NCE mean score of 53, which would indicate that they maintained their place in the distribution. This place or position represented by the NCE is relative to a greater sample (district or state) of test takers. The difference between the actual NCE and target NCE would be zero, if the cohort's mean fourth grade NCE was in fact 53. An NCE gain of zero would indicate that "expected growth" had been met. NCE gains greater than zero are termed "greater than expected growth," and NCE gains less than zero are termed "less than expected growth." When NCE gains are greater than zero and they exceed the standard error threshold, then it is increasingly likely that the growth is due to the teacher effect and school effect. In other words, teachers can be credited with producing greater than expected growth and "adding value" to their students.

Couched in these terms, VAM allows teachers and schools to be recognized for exceeding growth targets, even though achievement standards may not have been met. Rather than simply declaring that a group of students did not pass a particular state test, VAM can reveal that, in fact, those students made greater than expected growth and, as a result, are better positioned, thanks to their teacher's hard work, to approach the passing standard on the next year's state test. For these reasons, the proliferation of VAM, through TVAAS, into districts that had chronically underperformed and missed accountability standards signaled a new way to measure teaching contributions that may not have been reflected in student achievement data. As Braun (2005) pointed out, VAM introduced the promise of a much-need quantitative measure to gauge teacher effectiveness. The power to recognize settings (classrooms and schools) where underachieving students were routinely set on an upward trajectory was a capability previously unavailable. VAM also introduced the ability to reveal that a cohort of students who, by and large, might have passed a state test, according to the achievement standard, actually made less than expected growth, based on what their testing histories indicated about their potential. Again, beyond the confidence threshold of a standard error, that "less than expected growth" can also be ascribed to the teacher or teachers who taught those students. A narrow 'Passed/Did Not Pass' achievement-based view of student test performance could possibly mask such downward growth trajectories ("Battelle for Kids").

Beginning in 1993, value-added reports were furnished through the TVAAS system to schools and districts across the state of Tennessee showing progress measures for teachers of core tested courses in Grades 3-8 and selected high school math courses (Sanders & Horn, 1998). Protests immediately surfaced and multiplied in subsequent years. Chief among them was the contention that TVAAS did not account for socioeconomic levels or demographic factors, which are strongly tied to student performance on standardized assessments (“TVAAS Resources,” 2012). Initial replies from the Sanders camp asserted that TVAAS “circumvented” this problem; “Previous studies indicate that the influence of teachers and schools on the rate of gain are independent of the confounding socioeconomic factors” (Sanders & Horn, 1994, p. 309). While acknowledging that socioeconomic factors do influence attainment, or achievement, they claimed that VAM neutralizes the influence of those “confounding” factors, since students are only being compared to themselves. Several years later, Ballou, Sanders, and Wright (2004) would further claim that controlling for SES and demographic factors made very little difference to teacher effects estimated by TVAAS. These were merely the first volleys fired in a debate that would grow as the reach and influence of VAM expanded.

NCLB called for states to ensure that all of their teachers meet the standards of being highly qualified. The three criteria set out for teachers were: the attainment of a bachelor’s degree from an accredited institution of higher learning, full state teaching certification, and subject matter and teaching skills competency in the areas of their instruction (U.S. Congress, 2001). The logic, it seems, held that highly qualified teachers would deliver better instruction than less qualified teachers, and the results of that superior instruction would yield higher student achievement, as measured by standardized test scores, among other metrics. According to Phillips (2010), however, much of the literature on teacher quality has produced mixed results regarding the relationships between the three characteristics of a highly qualified teacher defined by NCLB and student achievement. With a parallel demand that schools and districts meet adequate yearly progress marks, NCLB itself began to shift the focus from *teachers* to *teaching* (Kennedy, 2006). If the metrics defining a highly qualified teacher yielded necessary, but insufficient conditions for student success, how could the actual teaching exchange be measured for effectiveness? Lasley, Siedentop, and Yinger (2006) noted that the rise of VAM in the 1990s coincided with the movement that came to define teacher education and preparation as a policy problem. As policies such as NCLB trained the spotlight ever brighter on teacher quality at a national level, VAM methodologies were ready and available to be adopted beyond the borders of Tennessee to fill the measurement void.

By the early 2000s, William Sanders had partnered with the SAS® Institute in North Carolina, a worldwide leader in applied statistics that offers analytical resources and consulting expertise to a wide range of industries. He declared that the increased testing requirements of NCLB would actually provide the robust infrastructure that would allow educators to better manage the progress of their students (Sanders, 2003). SAS® TVAAS morphed into EVAAS – the Education Value-Added Assessment System – and became the national leader in VAM applications. The EVAAS formulas were declared proprietary and SAS® began profiting handsomely from its adoption and use across the country. North Carolina and Ohio opened up new flanks in the use of VAM on either side of Tennessee and Kentucky respectively. An Ohio company filled a new market that was created by the demands of adopting EVAAS.

Battelle for Kids, based in Columbus, Ohio, was created in 2001 to provide school districts with a data interface that became a crucial piece of the VAM puzzle. In order for teachers and schools to receive accurate VAM data correctly attributed to educators, they need to accurately “link” to their students. In this process, teachers must verify each of their class rosters, report the percent of the school year that each student spent in their classes, and then specify the percentage of instruction that they delivered to those students in each course section. Battelle for Kids provides school districts with the tools (typically on a dedicated online portal) to complete the roster verification, or “linkage,” process. Districts using EVAAS then send their state test results along with their linkage output files to SAS® for analysis. Battelle for Kids’ work with school districts, while technically business arrangements, are characterized as educational partnerships in the name of student achievement.

With the rise of VAM from statistical curiosity to full-fledged product and service, conditions became ripe for the establishment of the Value-Added Research Center (VARC) in 2004 at the University of Wisconsin, Madison. Its dual tasks are to conduct research on the use and utility of VAM and work with school districts and states to set up these models (VARC website). VARC has even developed VAM systems that differ methodologically from EVAAS, but still deliver similar data on student progress measured by gains that can be attributed to teacher and school effects.

By the mid-2000s, the Department of Education opened competition for the Teacher Incentive Fund (TIF) grants. As the name suggests, most of the grant money delivered to grantee districts was to be paid out to teachers in recognition of their effectiveness in the form of incentives and rewards. But, how would teaching effectiveness be measured? Grant guidelines called for districts to reward teacher performance, as measured by newly adopted observation rubrics, and to reward student performance, as measured by student progress metrics, i.e. VAM (TIF website). The adoption of alternative teacher observation rubrics became a grant requirement due to the ubiquity of over-inflated teacher evaluation scores, which did not distinguish between varying levels of performance. Excellence went unrecognized and poor performance went unaddressed (Weisberg et al., 2009). New rubrics promising the re-framing of authentically capturing teacher classroom performance (inputs), VAM data characterizing student growth (outputs), and differentiated compensation were all serious departures from traditional practices. After the first three rounds of TIF grants, over 70 districts spread across 30 states and the District of Columbia were participating by 2010.

The Race to the Top (RTTT) Fund, steeped in the American Recovery and Reinvestment Act of 2009 (ARRA), placed over $4 billion out for competitive bids from states to improve public education. Among a host of required elements, RTTT applicants had to demonstrate how they intended to develop a longitudinal data system and use data to improve instruction. Applications that refused to implement VAM were downgraded in the competition process for RTTT funds (Yeh, 2012). That VAM now became a tacit requirement for federal monies marked not only its arrival, but also its entrenchment in the American education lexicon and landscape. The federal authorization of massive grants like TIF and RTTT for specified purposes has been seen by some as creating a resource-dependency relationship with states and districts (Malen, 2011). School systems have encountered extreme difficulty in their efforts to sustain piloted practices once grant periods conclude. The TIF incentive and reward budget amounts are not easily absorbed or replicated in state and local budgets once federal funds are removed.

Criticism of VAM

With the evolution, expansion, and widespread application of VAM, the bank of criticisms has exponentially grown, not the least of which is the question of how many "users" understand VAM. EVAAS is a highly sophisticated model and needs to be in order to work correctly, but this immediately places it at distance from its users (Amrein-Beardsley, 2008). With increasingly important consequences tied to VAM data, the need for it to be understood and used has never been greater. However, the system's opacity is worrisome to many educators, who have come to view it as some sort of magic that they will never understand. Ewing (2011) even contends that the statistical sophistication of VAM is used to intimidate educators and policy makers, which amounts to an aggressive misuse of mathematics. Many of the issues and protests surrounding VAM, beyond simply understanding it, range from the philosophical to the logistical to the ethical.

The first basic assumption, on which VAM has been built, is that standardized tests can accurately measure students' knowledge and skills at a given point in time and over time. This seems to be a tenuous foundational assumption, according to Amrein-Beardsley (2009) for one. While the question raised here can be applied well beyond the discussion of VAM, it's an assumption that all educators accept, for lack of a surer or better option. However, under the conditions of VAM, its importance is magnified when test scores are tied back to the teachers of students who produced them. This type of direct attribution, "assumes that student learning is measured well by a given test, is influenced by the teacher alone, and is independent from the growth of classmates and other aspects of the classroom context" (Darling-Hammond, Amrein-

Beardsley, Haertel, & Rothstein, 2012, p. 8). The conception of direct, isolated teacher effects is one that would be tough to reconcile with simple observations of educational settings. On this point, many have called for an acknowledgement that teacher effectiveness is not context free; rather, it is context specific (Newton, Darling-Hammond, Haertel, & Thomas, 2010). Even under the assurance that students are only being compared to themselves and the purported absence of SES and ethnicity effects, the question remains whether VAM systems can truly tease out the multitude of other factors that impact student achievement and growth.

Another basic assumption of the VAM process has been highlighted as a logical misstep. Kupermintz (2003) has questioned whether the interpretation of VAM data is based on a circular definition of teacher effectiveness. When individual students or cohorts of students are recognized as having made greater than expected growth, the effectiveness of their teachers is revealed. However, that teacher effectiveness is both defined and measured by student gains. This is not unlike declaring that a basketball team that scored more points than their opponents in a game is, in fact, a winning team.

One of the chief logistical concerns involves “spillover” and “residual” teacher effects. For example, while students’ test scores on secondary reading tests are influenced by the teacher quality of their English teachers, they also may be influenced by the quality of their mathematics teachers (Koedel, 2009). This would represent a case of spillover effects across subject areas. Sanders and Horn (1998) noted that the effects of an ineffective teacher continue to impact students in subsequent years, even when they are assigned to more effective teachers. These lingering residual effects, along with spillover effects, further cloud the issue of teacher attribution for VAM gains

(McCaffrey, Lockwood, Koretz, Louis, and Hamilton, 2004). In fact, spillover effects could be impacted by grade level and department team members in schools where collaboration is an expectation, or even by the climate set by the school administrative team (Briggs & Weeks, 2011).

Some teacher preparation programs have begun using VAM data from their recently placed graduates as another metric to assess the effectiveness of the program's quality. Again, disentangling the influence of effects is problematic. As Floden (2012) suggests, among other factors, the placement school effect could be greater than the preparation program effect on individual teachers' VAM data, particularly when they are placed in hard to staff schools and districts. While appealing in theory, it may prove patently unfair to judge entire teacher preparation programs on the VAM scores of their graduates when placement, assignment, and a host of other factors exert more influence on the data (Lincove, Osborne, Dillon & Mills, 2014; Plecki, Elfers, & Nakamura, 2012).

In some reward models constructed through the implementation of TIF or RTTT grants, teachers ranked in the top quartile based on district VAM data are the only ones that earn reward payouts. In this scenario where teachers are being compared to each other rather than a standard criterion, there are a limited number of "reward-able" places. Lockwood, Louis, and McCaffrey (2002) have noted that attempting to translate VAM gains into percentiles for the sake of ranking teachers is highly problematic, even under ideal statistical conditions. The forced distribution will also always count half of the teachers as below average (Scherrer, 2012). One such TIF district with rewards set up this way is the Houston Independent School District (HISD). Holloway et al. (2011)

found that some HISD teachers have become savvy with their teaching assignments and have initiated moves into grade levels and subject areas where it has been easier to generate VAM gains and earn reward payouts. One of the consequences of this is that talented teachers may be moving away from the subjects or grade levels they teach the best and enjoy the most (Hanushek & Rivkin, 2010b).

Another logistical issue focuses on the actual tests used as the raw material for VAM calculations. Some researchers have suggested that the selection of test can produce different results and identify different sets of teachers as being more effective than others (Papay, 2011). Even the subsections or constructs within a test can influence teacher estimates of effectiveness. In their examination of the VAM scores of middle school mathematics cohorts, Lockwood et al. (2007) found that, "value-added teacher effect estimates calculated from total scores may be sensitive to the relative contributions of each construct to the total scores" (p. 57). Some mathematics exams are made up of sub-tests, or constructs within a greater test, that may address, for example, procedural calculations and problem solving. By weighting one of these constructs over another, the relative ordering of the teacher effects may be subject to change.

As Hill, Kapitula, and Umland (2010) remind us, a fundamental assumption of VAM is that they should correlate or converge with expert ratings of instructional quality. The reality is that this is not always the case. In HISD over half of teachers reported that their EVAAS reports did not match their supervisors' observational scores (Amrein-Beardsley & Collins, 2012). If both metrics are designed to capture and characterize teacher effectiveness, shouldn't they do just that? Or is it okay that they don't align?

One unintended consequence of mismatches is that some supervisors might skew their observation scores to match the VAM scores.

Probably the most wide-ranging logistical issue centers on VAM participation. Nearly two-thirds of all teachers do not receive VAM data. Because VAM uses standardized test scores and compares those scores longitudinally, actionable data is only produced in reading, mathematics, science, and social studies from Grades 3 and above. Teachers in other grade levels and subject areas are not directly involved in the process and do not have VAM data attributed to them. This becomes a problem in reward models where a significant portion of payouts is based on VAM gains. Those teachers who do not “generate” value-added data are often relegated together to a sideline group that may or may not earn a reward based on how the school-wide VAM gains turn out – something they have little, if any, direct control over.

As the stakes rise, as reward money enters the picture, and employment decisions ride on VAM, criticisms are more often steeped in ethical terms. As Amrein-Beardsley (2009) underscores, using a single indicator to make consequential decisions violates long-held, fundamental standards set forth in the social sciences. For example, never would a student be diagnosed with a learning disability based on a single observation or a single work sample. Because VAM scores tend to be unstable over time, teachers may be rated effective one year and rewarded thusly, and then rated ineffective the next year, when their instructional practice might have been identical. Sass (2008) noted in a study comparing VAM rankings of individual teachers in consecutive years that nearly 15% of teachers ranked in the top quintile one year fell to the bottom quintile the next year and an equal proportion moved from the bottom to the top. Another recent longitudinal study using a 10-year sample of teacher VAM data showed variance within teachers from year to year with the greatest variance occurring at the lowest levels (Goldhaber & Hansen, 2012). One such teacher in HISD whose rating changed from effective to ineffective characterized this as a misuse of information; VAM was designed to measure students, he contended, not teachers (Banchero & Kesmodel, 2011).

Not all critics have been critical of VAM; some have taken on the challenge of brainstorming ways to constructively apply what VAM appears to reveal about teaching and learning. Harris (2009), in an article optimistically titled “Teacher Value-Added: Don’t End the Search Before It Starts,” takes a deep breath for the research community and proposes two new assumptions. First, VAM probably provides some useful information about teachers; second, this information is probably at least as useful and revealing to us as noting teacher credentials and conducting teacher observations. Working from those two starting points, perhaps VAM shouldn’t be rejected outright by its critics. Glazerman et al.. (2010) raise the ‘compared to what’ argument: What are the quantitative alternatives? VAM, flawed as it may be, they argue, is still a better determinant of student test scores than other measureable factors of teachers. Rivkin (2007) also cautions that even though VAM may not fully reveal the quality of classroom instruction, that doesn’t imply that it has no productive use.

Braun (2005) nudges the support further by proposing that, “VAM may ultimately offer a more defensible foundation for teacher evaluation than, say, methods based on absolute levels of student attainment or the proportion of students meeting a fixed standard of performance” (p. 4). By “defensible” Braun may be appealing to the quantitative-leaning side of researchers and educators in the spirit of “numbers don’t lie.” Teacher observations, after all, are qualitative events more open to influence and skew. We can even imagine the scenario where a school administrator might be swayed to assign unearned observation rubric scores at such a level that ensures a reward payout for a teacher that can use the extra money. At least VAM data are not open to such influences.

Another tack in constructively using VAM is placing it in context along with other measures of student achievement and teacher effectiveness (Prince, Koppich, Azar, Bhatt, & Witham, 2010). Recognizing its usefulness, but not assigning too much importance to it could be an effective compromise. Perhaps VAM initially took on too much importance because of its novelty. Di Carlo (2012) suggested that districts set minimum weights for VAM within an array of multiple measures to inform teacher effectiveness – 10%-20%, for example – and then experiment with adjusting those percentages. Whatever the weight given in overall assessments of teacher effectiveness, positive, stable VAM data over time might at least reveal settings and practices worth studying so that perhaps some of the most effective instructional practices might be applied generally (Ferrao, 2012). Meyer and Dokumaci (2009) also argue for the use of VAM to evaluate the effectiveness of instructional programs, and policies. Shifting the focus from individual teachers to wider systemic factors might prove more informative. Schools and districts can use the data to gauge whether their curriculum is consistently yielding student growth, for example.

Another constructive application of VAM involves aggregating the data within reward models. Reward models focusing on individual teachers and their VAM data are fraught with potentially explosive consequences. Teachers in departments or on grade levels that serve as mentors and do the bulk of the planning for the group may be the ones left out of rewards; their colleagues' rewarded VAM gains might have been produced as a result of those common plans and coaching sessions. An ongoing concern about using teacher-level VAM are the consistent research findings that students are not randomly assigned to classes with administrators, teachers, and parents all playing roles in the process (Paufler & Amrein-Beardsley, 2013). The nonrandom assignment of students to classes likely impacts or biases individual teachers' VAM scores. Jockeying for particular students may result, where some are seen as the ones most capable of exceeding growth targets. One levelheaded alternative to this is the proposal to build VAM rewards structures on the basis of teacher teams (Harris, 2010). Not only does it stabilize the data at the aggregate level by using greater sample sizes (Hanushek & Rivkin, 2010a), this approach could facilitate more cooperation, coordination, and collaboration among teachers. Teams either earn rewards all together or they do not; in the process all of the teachers on a team are responsible for all of the students in their particular grade level/subject area.

As a post-script to this discussion, during the year this research was conducted, the American Statistical Association (2014) issued a strong statement on using VAM for educational assessment, in which they sounded a host of cautionary notes. Chief among the concerns was the use of VAM data for high stakes personnel decisions such as hiring, firing, placement, and compensation. The quality of education, they claimed, is not one event, but a vast system of interacting components; placing too much emphasis on the data that one of those components yields is folly, if not outright irresponsible. The statement further noted that VAMs predict only future test performances, not necessarily broader learning outcomes. The subtext to that point is the question of whether tests themselves are sufficient indicators of long-range learning outcomes. The statement closed with a familiar statistical refrain, namely VAMs measure correlation, not causation; effects attributed to teachers may actually be caused by other factors not captured in the modeling.

The Framework for Teaching

Charlotte Danielson's framework for teaching (FFT) has become one of the most widely adopted teacher observation rubrics in the wake of grant requirements, such as TIF, and state policy changes, such as the Performance Evaluation Reform Act of 2010 in Illinois (Alvarez & Anderson-Ketchmark, 2011). It was designed, as the name clearly suggests, as a framework to guide conversations about teaching practice (Danielson, 2007). While the FFT has evolved into a rubric to evaluate teaching, its primary intent is to improve teaching by providing a common language to describe teaching practice. As Alvarez (2011) further noted, it was designed to be used with teachers across the continuum of experience, from novices to veteran practitioners.

The FFT is comprised of 22 components of teaching responsibility, which are organized into four domains: planning and preparation, classroom environment, instruction, and professional responsibilities.

Table 1

Framework for Teaching: Component Level View

Domain 1 Planning and Preparation	Domain 2 Classroom Environment
a. Demonstrating Knowledge of Content and Pedagogy	a. Creating an Environment of Respect and Rapport
b. Demonstrating Knowledge of Students	b. Establishing a Culture for Learning
c. Selecting Instructional Outcomes	c. Managing Classroom Procedures
d. Demonstrating Knowledge of Resources	d. Managing Student Behavior
e. Designing Coherent Instruction	e. Organizing Physical Space
f. Designing Student Assessment	
Domain 4 Professional Responsibilities	Domain 3 Instruction
a. Reflecting on Teaching	a. Communicating with Students
b. Maintaining Accurate Records	b. Using Questioning and Discussion Techniques
c. Communicating with Families	c. Engaging Students in Learning
d. Participating in a Professional Community	d. Using Assessment in Instruction
e. Growing and Developing Professionally	e. Demonstrating Flexibility and Responsiveness
f. Demonstrating Professionalism	

The layout of Table 1 is intentional to demonstrate two features of the framework (Danielson, 2007). First, it is intended to represent the cyclical process that leads from planning to instruction, and through reflection back to planning. Second, the two domains listed in the right column contain the components that are directly observable while students are present. That is, during a classroom observation evidence is gathered for the components in Domains 2 and 3 – classroom environment and instruction. Evidence can be gathered for Domains 1 and 4 – planning and preparation and professional responsibilities – through conferences and an examination of artifacts, such as lesson plans, assessments, and family contact logs. The four “levels of performance” used in Danielson’s framework are: unsatisfactory, basic, proficient, and distinguished. These levels of performance, along with their detailed descriptions of practice, according to Danielson, “permit the discussion about teaching to be non-personal; that is, if an evaluator cites events from a classroom observation as evidence for a certain placement on the levels of performance, the language serves to mediate the conversation” (Danielson & McGreal, 2000, p. 35).

Engaging students in learning, a component from Domain 3, is the centerpiece, or heart, of the entire FFT (Danielson, 2007). What is required for student engagement, Danielson explains, is intellectual involvement. Without this crucial piece in place, a lesson stands little chance of succeeding. Often in elementary mathematics lessons, for example, the mere presence of hands-on manipulatives is thought to be sufficient to engage students. Danielson cautions, “Instructional materials and resources are not, in themselves, engaging or unengaging; rather, it is a teacher’s and student’s use of the materials that is the determinant” (Danielson, 2007, p. 84).

Both Danielson and Phillip Schlechty write at length about what is and isn’t student engagement: working busily may not indicate engagement, yet it is often mistaken for it. Schlechty (2011) identifies five levels of engagement, and only really one of them qualifies as the highest form of “minds-on” engagement, as Danielson would term it. The second and third levels are called strategic compliance and ritual compliance. They are characterized by behavior that is within expectations; completed work handed in on-time, cooperative behavior, but minimal interest and potential for retention. Strategically compliant students will put forth maximum effort and complete all assignments to the best of their abilities, but may not truly engage with the content of lessons. Ritually compliant students will go through motions and put forth minimal efforts to get by. The fourth level is called retreatism, which is characterized by students withdrawing from the instructional exchange, not participating, and just hoping to go unnoticed. The final stage is rebellion – these students will either occupy themselves quietly with something entirely off-task, or they will openly disrupt instruction as they try to extract themselves from the instructional exchange. The goal then for teachers is to move students toward the top two levels of engagement where their best efforts are consistently produced, always being mindful, however, of merely compliant behavior.

The FFT is both praised and criticized for one of its chief features – its generic applicability to a wide variety of educational subjects and settings (Kimball, White, Milanowski, & Borman, 2004). Its advantages are many. Administrators can apply a common vocabulary to their discussions with teachers about instructional practice, regardless of content area or grade level. However, sometimes lost within the parameters of that common language is the vocabulary to make detailed insights and provide specific feedback on instructional practices that are very much content-specific. One of the drawbacks, however, of using content-specific rubrics is the sheer number that school administrators would need to master and use, and the tendency for the rubrics to become unnecessarily detailed, unwieldy, and labyrinthine; clarity is always preferable (Schmoker, 2012).

Mathematics Pedagogy

While disciplinary knowledge, or deep familiarity with a subject, appears to be a prerequisite for effective instruction, it is not enough. “Teachers use pedagogical techniques particular to the different disciplines to help convey information and teach skills. General pedagogical skill is insufficient; every discipline has its own approaches to instruction” (Danielson, 2007, p. 45). Hill, Rowan, and Ball (2005) studied mathematics teachers’ mathematical knowledge for teaching (MKT) to gauge their facility in applying content specific pedagogy to their mathematics lessons. This project separated teachers’ knowledge of mathematics from their ability to communicate mathematical ideas using appropriate materials, suitable representations, and terms that are both accurate and comprehensible to the level of their students. Participating teachers were given a test that measured their MKT as high, medium, or low. High MKT scores were positively correlated with student achievement in a study of school improvement (Hill et al., 2008).

After determining that MKT and student achievement might correlate, Hill et al. (2008) further developed a mathematical quality of instruction (MQI) rubric to analyze classroom practice. There are six main rubric elements: mathematics errors, responding to students inappropriately, connecting classroom practice to mathematics, richness of the mathematics, responding to students appropriately, and mathematical language. The first two elements, which are phrased in the negative, track errors that teachers make in language and representation, and misinterpretations and misunderstandings teachers make that might reinforce misconceptions. The middle two elements focus on the rich integration of mathematical thinking and representation into lessons, which might develop such skills as justification and reasoning, all the while granting increased access to mathematical ideas. The final two elements acknowledge when teachers correctly interpret students' utterances and questions, and respond with clear, accurate information that might help develop new and correct conceptions.

Capraro, Capraro, Carter, and Harbaugh (2010) state that teachers' use of representational models provide insight into teaching quality because the representational forms modeled by teachers have tremendous influence on how students develop their own mathematical understandings. When the representations and modeling are deliberate, accurate, and comprehensible, students are better positioned to construct understanding. This is the ultimate task of mathematics lessons – to facilitate students in developing their own understanding. It's not enough to understand mathematics, effective mathematics teachers need to know how to communicate and teach mathematics so their students can make sense of the concepts and apply the necessary skills. This involves, "the bifocal capacity to understand ideas and to see them from the perspectives of others who are first encountering them" (Ball & Forzani, 2010, p. 10).

The construction of mathematical tasks must, therefore, require students to think, reason, and make sense of mathematical ideas (Boston, 2012). Chapin and Johnson (2006) urge teachers at every turn in their book *Math Matters* to "emphasize sense making in all mathematical activities" (p. 131). This conceptual approach to mathematics instruction (as opposed to a procedural one) develops even greater understanding, they contend. Burns (2000) asserts that understanding why is as important as knowing how; both are necessary in mathematics. For example, knowing how to figure percents is not sufficient for choosing the best money market account. "These decisions," Burns explains, "require understanding and judgments that extend beyond algorithmic thinking" (p. 151).

Correlating VAM and Observation Data

Several studies have been conducted to find correlations between student achievement gains and teacher observation rubric scores. Stronge, Ward, & Grant (2011) studied over 300 fifth grade teachers and examined their value-added gains alongside their observation scores, which were based on an eclectic, hybrid rubric. Teachers who scored in the top quartile in VAM gains scored significantly higher ($p < .05$) than the other teachers in rubric dimensions related to classroom management, classroom routines, organization, and availability of appropriate instructional materials. They also scored higher in establishing a culture of respect and rapport indicating that the components within the FFT domain of classroom environment are crucial, perhaps even more so than the components in the domain of Instruction.

A study by the Measuring Effective Teaching (MET) Project (Kane & Staiger, 2012) in which they examined the FFT, among several other rubrics, and compared scores earned on those instruments to student achievement gains (working in the reverse direction of the Stronge, Ward, and Grant study). Students with teachers who scored in the top quartile of observation scores moved ahead of comparable peers by a minimum of 1.5 months, according to test gains. Students with teachers who scored in the bottom quartile of observation rubric scores fell behind comparable peers by a minimum of 1 month, according to test gains (Kane & Staiger, 2012, p. 8). One current emerging in all of these studies, and articulated in this MET study is the idea that multiple measures of teaching effectiveness might lead not only to higher predictive power, but also to greater reliability.

Hill, Kapitula, and Umland (2010) assert that VAM data should converge with expert ratings of instructional quality and estimates of teachers' knowledge (e.g. MKT), and *not* converge with unrelated constructs, such as the population of students in a given classroom (p. 799). A second caution comes from Milanowski (2004), whose study of the Cincinnati Public Schools was one of the earliest attempts to find correlations between teacher and student performances. He urges us to consider two issues regarding teacher observations: first, how well do the components in a rubric correspond to our conception of teacher behavior that ought to contribute to student learning, and second, how well do the judgments of observers actually capture the teacher behaviors described in the rubric. These are fundamental questions that need to be addressed in any serious effort to use rubrics to examine and improve practice.

The Observation Process

The FFT was designed to provide a lens and language through which to provide meaningful feedback to teachers on their instructional practice; all FFT observations are considered formative events. The effectiveness of formative teacher observations, however, according to Looney (2011), depends largely on how feedback is given and whether teachers have the opportunities to discuss their practice on a regular basis. School and districts adopting the FFT frequently incorporate structured pre- and post-conferences in order to provide teachers the forum to discuss their practice with administrators. The structure of the conference-observation-conference cycle, even for formative purposes, often necessitates scheduled observation times. Marshall (2009), however, claims that teacher performances during scheduled observations are often stilted, the observer effect is a factor, and extensive write-ups are difficult, time-consuming, and seldom yield useful, actionable information. Rather, more frequent and less formal observations with accompanying feedback provides administrators with more representative evidence and teachers with more authentic feedback.

Walkthrough observations – quick, unscheduled, less formal, more frequent classroom visits with less-structured feedback – are typically employed as supplements to more formal observations (formative or summative) within many observation systems. To further delineate the purpose of each classroom visit, some walkthroughs might be designated as non-evaluative, conducted by instructional coaches, and others might be designated as part of an overall evaluation, conducted by administrators (Milanowski, 2011). Additionally, while formative and summative evaluations have different purposes – to coach and to judge, respectively – many doubt whether they can effectively be conducted in exclusion from each other's influences (Towndrow & Tan, 2009; Milanowski, 2005).

Logistics often impede the smooth implementation of the observation process at many schools, regardless of the design or purpose of evaluations. The daily demands of school management often conspire to make many principals harried, unfocused, and, as a result, they may not take the time to confront bad teaching (Marshall, 2009). Additionally, as Kimball and Milanowski (2009) note, a school administrator, "who views the performance evaluation system as too much work or just another mandate is likely to spend less time observing teaching behavior and making careful assessments" (p. 39). Firestone (2014) has openly wondered if there are enough skilled administrators to take on the dual responsibilities of providing truly useful feedback to teachers *and* doing all of the other things necessary to run a school.

Despite the daily challenges, Rockoff and Speroni (2010) have emphasized that in order for evaluation systems to perform optimally, the issue of consistent implementation and uniform application of standards by all trained evaluators should be addressed. Thorough training of observers, calibration activities to establish inter-rater reliability, and explicit observation guidelines can all help strengthen observation processes. Teacher trust in the process can be grown when they know that all observers are equally prepared, ensuring that their scores are dependent on the strength of the observation instrument, not on the arbitrary assignment of observer. To this point, Cantrell and Scantlebury (2011) emphasize, for teachers “it is patently unfair for their rating to be dependent upon the ability of the rater rather than the quality of the lesson” (p. 31). Another way to strengthen an observation process is to introduce regular observations conducted by outside observers trained in the use of the rubric. According to Whitehurst, Chingos, and Lindquist (2014), observations conducted by outside observers are, in many cases, more valid because of their perceived objectivity. Incorporating outside observers with little prior knowledge of the teacher, the specific classroom context, or the school can actually strengthen the processes of gathering evidence and examining that evidence through the lens of the rubric.

Beyond the opportunities to provide formative feedback to teachers and craft summative annual evaluations, the observation process has other uses. Matula (2011) states that an essential function of any teacher evaluation system is the identification of low-performing teachers for potential removal. After diligent efforts to improve a teacher’s practice, the responsible application of observations/evaluations would be to recommend the non-renewal of contracts for teachers not showing improvement from sub-par levels of instructional practice. However, Matula emphasizes that the process “must not only *be* fair, it must be ‘perceived’ by all teachers as being fair (p. 119). It makes instructional sense, particularly for struggling students who cannot afford to fall farther behind while their ineffective founders. Further, it makes economic sense to move out low-performing teachers; salary increases can be better justified when they apply to all teachers who have demonstrated, minimally, that they are effective enough to continue in a school system (Hanushek, 2011).

While a teacher observation/evaluation system might help improve instructional practices through intentional feedback cycles, identify effective practitioners, and influence retention decisions, it is only one instrument in a vast array of factors that continually shape education. According to Akiba and Letendre (2009), an effective teacher evaluation process is just one factor in a vast system to improve teaching; we can’t evaluate our way to developing a high-quality workforce. Other key components include recruitment, hiring, induction, placement, working conditions, and ongoing professional development.

CHAPTER 3

METHODOLOGY

The purpose of this study was to investigate the extent to which teacher performance and student performance measures correlated, and to understand which specific practices of highly rated mathematics teachers might have related to increased student performance.

The following two research questions guided the study:

1. What is the relationship between students' progress scores in mathematics, as measured by VAM, and their teachers' instructional performance scores, as measured by the FFT?
2. What is the relationship between levels of teacher and student performance as articulated in interviews with campus administrators, and an examination of their documented observation evidence using the FFT?

These were the working hypotheses for this study:

1. It was hypothesized that there would be a significant relationship at the $p < .05$ level between teacher performance FFT scores and student performance VAM scores.
2. It was assumed that student value-added progress measures would positively relate to teacher instructional performance ratings. Specifically, it was assumed that teachers rated high according to their FFT observation scores would also be rated high according to their VAM student progress scores, and teachers rated low on FFT would also be rated low on VAM. In cases where the ratings do not match, any number of factors might account for the mismatch, including the

possibility that one of the underlying suppositions might be flawed, namely that these are, in fact, both measures of effective teaching. It was also assumed that interviews with school administrators and an examination of their observation documentation would provide insight into confirmatory or discrepant VAM/FFT relationships. Finally, it was hoped that this study might yield a set of best teaching practices by examining the most highly rated instructional practices of those teachers who scored high on both measures.

Setting

This research was conducted at five elementary schools in a large urban school district in north Texas. These five schools are part of a 14-school cohort of campuses participating in the district's federal Teacher Incentive Fund (TIF) grant. The cohort is comprised of Title I schools that have been historically hard to staff and under-performing. Most students (> 90%) attending these schools are either African-American or Hispanic and qualify for free or reduced lunch (> 85%). This study was conducted during the third year of TIF grant implementation. The main features of grant implementation have included the adoption of an alternative teacher observation rubric, the Charlotte Danielson framework for teaching (FFT), multiple teacher observations conducted by multiple observers, a performance based compensation system (PBCS) that has included incentives and performance rewards for teachers, and the use of value-added modeling (VAM) to measure student progress.

As part of the district's commitment to improving this cohort of schools, the school board authorized local funding to enhance grant implementation for the purpose of hiring extra personnel at these campuses. Each of the five elementary schools has three additional teaching assistants, a data analyst, and a teacher on special assignment to be used in a variety of instructional capacities. Each of the nine secondary schools in the cohort has a dean of instruction and a data analyst. Every campus in the cohort articulates an annual rededication commitment agreement (RCA), comprised of the school's goals, priorities, and expectations for the upcoming year. By signing the RCA, teachers and administrators re-commit to the ongoing project of improving the school. One of the incentive payments in the PBCS is based on fulfilling and upholding the RCA. With the introduction of the FFT as an alternative teacher observation instrument to the state required Professional Development and Appraisal System (PDAS), the district school board granted a waiver to these schools, which allowed less-than-annual PDAS observations for non-probationary, continuing teachers in good standing. Additionally, these schools have been grouped together in one of the district's three learning networks under the direction of the same set of leadership directors in order to streamline the administration of the grant and provide the campuses with operational flexibility within the district.

Population

To maintain participant anonymity, codes were assigned to all participants. The five campuses were labeled A through E. Mathematics teachers were assigned numbers per campus, e.g. AMT1 (Campus A, Mathematics Teacher 1). Administrators were also assigned numbers per campus, e.g. AAD1 referred to Administrator 1 at Campus A.

Teachers

There were 28 teachers, whose data – ID&E scores and VAM scores – were used in this study. They were indirect participants and were not contacted about this study. Since district elementary schools have the flexibility to departmentalize upper grades, of the 28 participating teachers, 5 were the sole mathematics teacher in the grade level. These five teachers taught all grade level mathematics sections. The other 23 teachers were part of a team of grade level mathematics teachers; teams ranged in size from 2 to 4 teachers. In most cases, these 23 teachers were self-contained and responsible for instruction in all subjects for one class of students, meaning they taught only one section of mathematics. Several taught more than one section, but due to the size of the grade level, they were still not the sole mathematics instructor.

Campus Administrators

Both administrators – the principal and assistant principal – at each of the five TIF elementary schools participated in this study. After accepting an invitation to participate in the study, each administrator signed an informed consent form, in which the scope of the study and their participation were explained. The ten administrators were responsible for observing and evaluating teachers at the campuses twice during the year using the FFT. Teachers were observed in the fall semester by one of the campus administrators and in the spring semester by the other campus administrator. The ten administrators were direct participants in the study and were interviewed about the FFT observations they conducted of the mathematics teachers in Grades 3-5. In the interviews participants discussed the evidence gathered, the component scores administered, and their perceptions of effective mathematics instruction.

Table 2 summarizes the administrators' experience including how many years they had held their current position and how many years they had been involved with the TIF grant using the FFT as an observation instrument. Additionally, Table 2 shows whether or not each administrator had previous mathematics teaching experience in the primary elementary grades (Pre-K - Grade 2) or in the upper elementary grades (Grades 3-5), and whether or not the administrator had previously served as an elementary mathematics instructional coach within the district. The mathematics coach position was campus-based and was designed to provide ongoing teacher support, co-teaching partnerships, and targeted professional development in mathematics instruction to all grade levels. Three of the ten administrators – AAD2, BAD2, and EAD2 – previously served in this capacity. As such, all three gained extensive experience co-teaching and coaching mathematics in Grades 3-5.

Previous teaching experience varied across the ten administrators. Of the three former coaches, both BAD2 and EAD2 also had previous experience teaching mathematics in Grades 3-5; AAD2 did not have elementary teaching experience, but taught mathematics in the middle school grades. Three of the remaining seven administrators had experience teaching mathematics in Grades 3-5: DAD1, DAD2, and EAD1. The other four administrators – AAD1, BAD1, CAD1, and CAD2 – had experience teaching mathematics at the primary level only. Three of the administrators had been in their current position over ten years, while the other seven had five or less years experience in their current position. Five of the administrators were in their third year of using the FFT, two were in their second year, and three were using it for the first time during the year of this study.

Table 2

Administrator Experience

Administrator	Yrs. Current Position	Yrs. FFT Experience	PK-2 MT Experience	3-5 MT Experience	Elem. MC Experience
AAD1	17	1	Y	N	N
AAD2	1	1	N	N	Y
BAD1	10	3	Y	N	N
BAD2	4	3	N	Y	Y
CAD1	2	3	Y	N	N
CAD2	4	2	Y	N	N
DAD1	2	2	N	Y	N
DAD2	5	1	N	Y	N
EAD1	13	3	Y	Y	N
EAD2	4	3	N	Y	Y

Sampling

The sampling for this study was a non-random, convenience sample. All ten of the administrators from the five campuses participated in the study. Additionally, the FFT and VAM data from all 28 mathematics teachers in Grades 3-5 from the five campuses were studied.

Design

This was a mixed methods study. The FFT observations yielded both qualitative (written evidence) and quantitative (scores) data sources. Interviews conducted with the administrators about their FFT observations were another qualitative data source. The VAM student progress scores were a quantitative source that revealed the mean progress of each grade level cohort of students based on their prior testing histories. According to the typology of research designs explained by Teddlie and Tashakkori (2006), this was, more specifically, a sequential mixed methods study. These types of studies “answer exploratory and confirmatory questions chronologically in a pre-specified order” (p. 22). The order for this study was determined by the timeline for data availability: all teacher observations preceded the availability of student VAM data. The analysis of observation data explored which teaching performances were judged to be most effective, and interviews with the observers served to further elucidate those judgments. The analysis of the VAM data served to determine the growth of students’ from one year to the next. The correlation between the VAM growth scores and teachers’ FFT showed the relationship between teacher effectiveness and student growth.

Data Sources

ID&E Scorecard

While the FFT is composed of 22 components organized into its four domains, the district decided to use only a subset of 13 components in its TIF grant-mandated Individual Development and Evaluation (ID&E) scorecard. Ten components from Domain 2 ‘The classroom environment’ and Domain 3 ‘instruction’ were included in the

ID&E scorecard because these are directly observable in the classroom. Administrators gathered evidence during the classroom ID&E observation, which was sorted and coded into these ten components. The other three FFT components that were included in the ID&E scorecard provided bookends to the observation. Evidence was gathered during the mandatory pre and post conferences. The pre-conference discussion provided evidence for Component 1e ‘designing coherent instruction’ and the post-conference discussion provided evidence for Components 4a ‘reflecting on teaching’ and 4b ‘maintaining accurate records.’

The Danielson FFT has four levels of performance per rubric component: unsatisfactory, basic, proficient, and distinguished. For the sake of compiling overall scores and averages, that determine reward payouts, numeric scores were assigned to each level of performance. Since the unsatisfactory level is characterized by practice that is severely lacking, it was assigned zero points. The basic level was assigned two points, the proficient level three, and exemplary four. An average score was produced at the end of each semester per teacher. When both semesters’ observation averages were averaged again at the end of the school year, reward payout amounts were determined.

Administrator Interviews

Each of the ten campus administrators participating in the study was interviewed regarding the FFT observations they conducted of their mathematics teachers in Grades 3-5. The interviews focused on the evidence the administrators gathered for each of the rubric components and how they arrived at the scores they assigned. The goal of these interviews was to gain insight on what the administrators considered effective mathematics instruction and the extent to which their observations confirmed those beliefs.

VAM Data

VAM data were calculated for the district by the Statistical Analysis System (SAS®) Institute of North Carolina – a recognized leader in applied statistics – using its Educator Value Added Analysis System (EVAAS) methodology. Student scores on the State of Texas Assessment of Academic Readiness (STAAR) mathematics test were combined with “linkage” files of teacher-verified class rosters in the process. Mean normal curve equivalent (NCE) scores for each grade level cohort of students were compared with the mean NCE scores of that same cohort on the previous year’s test to yield a NCE gain, which indicated the extent of growth. NCE gains were then divided by the standard error of the calculation to reveal the cohort index, or NCE gain score. In cases where the index score was greater than 1.0, meaning the positive NCE gain was greater than 1.0 times the standard error, these were considered “greater than expected growth.”

The grade level cohort NCE gain score is the only level of VAM data currently released by the district where this study took place. District leadership has not authorized the release of teacher-level VAM data. For this study, the aggregated grade level cohort NCE gain score was assigned to each teacher on that grade level team. As mentioned above, five of the fifteen grade levels studied had a single mathematics teacher who taught multiple sections of mathematics. The grade level cohort NCE gain score, which was aggregated across all mathematics sections for each cohort, functioned as a de facto teacher-level score for those five teachers. The other 23 teachers were also assigned the grade level cohort score, even though their individual attribution was not defined.

Data Collection

ID&E Scorecards

Component scores*.* Teachers were observed once per semester by either the campus principal or assistant principal. The FFT was used for each observation for a minimum of 30 minutes. The observation was preceded within five school days by a pre-conference with the observer, and it was followed within ten school days by a post-conference. The pre-conference form (Appendix A) contained five questions about the scheduled lesson, and the post-conference form (Appendix B) contained four questions about the actual lesson delivered. Teachers were expected to complete the conference forms prior to the pre- and post-conferences, where their responses served as the basis for discussion. At the post-conference, the observer also provided the teacher with a draft copy of the ID&E scorecard (Appendix C) with a numerical score and evidentiary support for 12 of the 13 rubric components (minus Component 4a ‘reflecting on teaching’, which was evaluated during the post-conference discussion). The draft ID&E scorecard was discussed at the post-conference as well. A finalized and complete copy of the ID&E scorecard with scores and evidence was provided to teachers within several days after the post-conference. The completed ID&E scorecards, along with all other district walkthrough and PDAS observations, were stored and shared in the district Eduphoria electronic portal where teachers had password-protected accounts. Both completed conference forms were also uploaded and attached to the completed ID&E scorecard in Eduphoria.

The ID&E scorecard template in Eduphoria was structured to provide administrators with space to list their observation evidence for each rubric component beneath a description of the component. Administrators then assigned a level of performance for that component based on their listed evidence. To determine the level of performance, administrators consulted the full ID&E rubric (Appendix D), which described each of the four levels of performance in detail per component. Each campus team of administrators used a Google Docs spreadsheet to record, per observation, the following information: the name of the teacher observed, the name of the administrator who conducted the observation, the observation date, and the 13 individual component scores, copied from the ID&E scorecard. The spreadsheet was formulated to calculate an average score for the entire observation based on the 13 individual component scores. Further, the spreadsheet calculated cumulative averages for each rubric component at the campus based on all of the scores entered. Each teacher was observed twice during the school year, once per semester, and each of the two campus administrators conducted one ID&E observation per teacher since the TIF grant called for, not only multiple observations, but also multiple observers. The ID&E scorecards and the campus Google Docs spreadsheets were obtained.

Rubric evidence. As described above, administrators entered the evidence they gathered during classroom observations and the pre and post conferences with teachers into the ID&E scorecard template in Eduphoria. District training on the FFT process called for administrators to take notes during each classroom observation on what they saw and heard. After the observation, they were expected to sort and code their evidence, copying specific notes and statements into the ID&E template under the relevant component. It was not unusual for one specific evidence statement to be coded into more than one component and copied in both places of the template as support

Administrator Interviews

Each of the ten administrators was interviewed to discuss their thoughts on effective mathematics instruction, what they typically looked for in mathematics classrooms as signals of effective instruction, and what they had observed during recently completed ID&E observations of mathematics teachers in Grades 3-5. Eight of the ten administrators were interviewed during the month between the closing of the fall semester ID&E observation period and the winter holiday break. Six of the eight were interviewed individually; logistics necessitated a joint interview with the two administrators at campus D. The other two administrators, AAD2 and BAD2, were interviewed during the month between the closing of the spring semester ID&E observation period and the end of the school year. They were not interviewed in the fall because they had not completed any ID&E observations of mathematics teachers in Grades 3-5 during the fall semester. AAD2 and BAD2 conducted ID&E observations of mathematics teachers in Grades 3-5 during the spring semester only.

The interviews were semi-structured and focused on five basic questions (Appendix E). Depending on individual responses, follow-up questions were asked so that the administrators could explain in greater detail the reasoning used for score assignments. The interviews were recorded on an iPhone, using the app Evernote, transferred to an iCloud account and accessed through iTunes. Each interview was transcribed after which the transcribed data were loaded into the NVivo qualitative data analysis program. The data were sorted, coded, and analyzed applying an analytical approach whereby the five questions/topics served as categories. Data were coded by individual response within each category. An analysis of the data within each category was done for emergent patterns and themes.

VAM Data

The VAM analysis was conducted by SAS®, using the EVAAS calculation methodology. The district submitted the roster verification linkage file produced in partnership with Battelle for Kids along with the full district state test results file to SAS®, the VAM analysis company. The value-added data analyses were returned to the district in August, after which the subject level/cohort data for the five participating elementary schools were extracted from the district data set. For this study, only grade level mathematics data were reviewed. Teacher-level VAM data are not included in the district data; hence data reviewed encompassed grade level data, rather than individual student by teacher data.

The district test results included data from the STAAR test and the Stanford 10 test, which was administered in the primary grades (1st and 2nd) and in non-STAAR content areas, such as 4th grade science. Stanford 10 is typically given mid-year, the administration of the test is not as structured and controlled as with STAAR administration, and the state instructional standards (and the district curriculum) are not fully reflected in the content of the test. Third grade VAM results, which use Stanford 10 data as their antecedents, possibly reveal disparities between the Stanford 10 and STAAR in terms of test administration conditions and level of focus on state standards impacts measures of growth. A drop in mean NCE gain seen in student cohorts between second and third grade may not be surprising when the level of test rigor and the pressure of testing conditions increases.

Data Collection Timelines

The ID&E observation window for the fall semester closed in mid November. All completed ID&E scorecards were posted to Eduphoria, with the accompanying completed conference forms, by mid-December. The scorecards and the first semester component scores and teacher averages on the Google Docs spreadsheets were accessed in January. The spring semester ID&E window closed at the end of March. All completed ID&E scorecards and conference forms were posted to Eduphoria by the middle of April. The scorecards and the Google Docs spreadsheets, which included both sets of scores and averages for every campus teacher at the five schools for the 28 teachers were accessed in May.

Eight of the ten administrators were interviewed in seven sessions (DAD1 and DAD2 met together because of scheduling issues) in December. One of the interviews was conducted after school; the others were conducted during school hours at the request of the administrators. All interviews were held at the campuses in the administrators' offices. The final two interviews were conducted in May. One was conducted after school, one was conducted during school hours, and both were held at the campuses in the administrators' offices. The duration of the interviews ranged from 12 minutes to 47 minutes.

The STAAR mathematics tests were administered in April for all students in Grades 3-5. The linkage process was conducted in May, whereby teachers verified their class rosters. The district test results file and the linkage output file were sent to

SAS® in June. The VAM data were returned to the district and accessed for this study in August. At that time the comparative analysis of VAM and ID&E data was conducted.

Data Analysis

ID&E Scorecards

Component scores. The 13 component scores were averaged for each observation. Both observation averages per teacher were averaged again to provide an overall teacher ID&E score for the year. Since each of the five campuses yielded a different mean for overall grade level teacher scores, individual teacher ID&E scores could not be compared across campuses. To allow for a comparison of the 28 teachers' scores, the ID&E were converted to ordinal scores based on how individual teacher's scores compared to the campus mean scores. Four teaching performance (T) levels were established as shown in Table 3 below.

Table 3

Explanation of Teaching Performance Levels

Teaching Performance Level	Parameters
Level 4 (T4)	Teacher ID&E Score More than 1 SD above campus mean
Level 3 (T3)	Teacher ID&E Score Within 1 SD above campus mean
Level 2 (T2)	Teacher ID&E Score Within 1 SD below campus mean
Level 1 (T1)	Teacher ID&E Score More than 1 SD below campus mean

The rationale for converting scores into T levels determined by campus means and standard deviations was to account for variation in observer scoring tendencies across the five campuses. For example, if one particular campus pair of administrators tended to score harder than any other campus, which would result in an overall lower campus mean score, teachers they considered especially strong could still be identified as such. At campuses with a higher overall mean score, teachers identified as less effective could still be marked at T level 2, for example, even though their personal overall mean would yield a higher T level rating when compared to other teachers across the cohort.

Rubric evidence. Evidence statements provided in the ID&E scorecards for each FFT component were analyzed and compared to the explanations provided by administrators in their interviews. Administrators were expected to cite evidence from their classroom observations and list statements under each rubric component in the ID&E scorecard. The depth and specificity of these evidence statements revealed the basis for the administrators' judgments indicated in the levels of performance reported.

Administrator Interviews

Interviews were recorded, transcribed, and coded into categories. I transcribed all of the interviews myself in order to attend to the phrasing, hesitations, and intonations in the verbal responses that are not typically captured in the text of transcriptions that might be produced by a third party. Following grounded theory, which, as Glesne (2011) points out, is not a theory, so much as it is a methodology, emergent patterns and themes within the interview data were noted. NVIVO software was used to help organize and analyze the interview responses. Codes and concepts were extracted from the interview responses and applied across all of the transcripts to find commonalities with the goal of possibly positing over-arching themes. In a number of cases, administrators did not use content-specific terminology. This was an instance when my background as a mathematics content specialist and coach helped to interpret responses that were not expressed in terms common to mathematics pedagogy.

The goal of this process was to gain detail and insight into the most relevant evidence that the administrators used as the basis for the FFT observation scores they assigned to the rubric components. Administrators were asked to explain what they generally considered to be the characteristics of effective mathematics instruction and also to reflect on the ID&E observations they conducted and describe teachers they rated most and least effective. They were also asked about the FFT components they considered to be essential to effective mathematics instruction.

VAM Data

Similar to the ID&E teacher scores, the VAM cohort scores were converted into ordinal data. Three value-added (V) levels were established as shown in Table 4 below. These levels were based on how the mean NCE gains compared to the standard error for each cohort calculation. The index score (mean NCE gain divided by the standard error) showed whether or not expected growth, greater than expected growth, or less than expected growth was reached. An index score between 1.0 and -1.0 was considered expected growth; the NCE gain was within one standard error above or below zero. An index score greater than 1.0 was considered greater than expected growth, and an index score less than -1.0 was considered less than expected growth.

Table 4

Explanation of Value-Added Levels

Value-Added Level	Parameters
Level 3 (V3)	NCE Gain More than 1 SE above 0
Level 2 (V2)	NCE Gain Between 1 SE above/below 0
Level 1 (V1)	NCE Gain More than 1 SE below 0

Relationships

The two quantitative data sources (T levels and V levels) were combined to form a matrix for the placement of the 28 teachers based on their levels in each measure. Those teachers whose T level was high (Level 3 or 4) and whose V level was high (Level 3) were considered high T/high V overall. They are indicated in the table as T3V3 and T4V3. Teachers whose T level was low (Level 1 or 2) and whose V level was low (Level 1) were considered low T/low V overall. They are included in the table as T1V1 and T2V1. Teachers rated T3V1 or T4V1 would have mismatched measures with high T levels combined with a low V level. The other possible mismatch would occur where teachers might be rated T1V3 or T2V3, indicating a low T level combined with a high V level. All other ratings involve a V level of 2, indicating expected VAM growth, combined with either high or low T levels. These combinations are shown in Table 5.

Table 5

Matrix of Performance Measures

	Level V1 NCE Gain More than 1 SE below 0	Level V2 NCE Gain Between 1 SE above/below 0	Level V3 NCE Gain More than 1 SE above 0
Level T4 Teacher ID&E Score More than 1 SD above campus mean	T4V1	T4V2	T4V3
Level T3 Teacher ID&E Score Within 1 SD above campus mean	T3V1	T3V2	T3V3
Level T2 Teacher ID&E Score Within 1 SD below campus mean	T2V1	T2V2	T2V3
Level T1 Teacher ID&E Score More than 1 SD below campus mean	T1V1	T1V2	T1V3

The qualitative data sources were also combined and examined in relation to each other. Observation evidence cited in the ID&E scorecards was analyzed alongside the administrator interview transcripts. When administrators cited specific practices of specific teachers as either effective or ineffective during an interview, the corresponding components from the ID&E scorecard were studied in order to confirm those explanations and to understand the level of feedback provided to teachers about their practice. The goal of this part of the study was to understand why administrators rated their teachers as they did. In cases where ID&E observation evidence may have been scant or general, whether or not the corresponding scores were high or low, the interviews helped to provide more details and a deeper level of understanding behind those evaluative judgments.

Chapter Summary

In this chapter, the research methods and design of this study were explained. Specifically, this chapter included information about the participants, the data sources, processes and timelines for data collection, and the data analysis. In Chapter 4, the results of this study are discussed in addition to how the results relate to the original two research questions.

CHAPTER 4

RESULTS

The purpose of this study was to investigate the extent to which teacher performance and student performance measures correlated, and to understand which specific practices of mathematics teachers related to student performance.

The following two research questions guided the study:

1. What is the relationship between students' progress scores in mathematics, as measured by VAM, and their teachers' instructional performance scores, as measured by the FFT?
2. What is the relationship between levels of teacher and student performance, as articulated in interviews with campus administrators, and an examination of their documented observation evidence using the FFT?

In this chapter, the results are discussed by data source: ID&E component scores and recorded evidence, administrator interviews, VAM scores, and correlation between VAM and ID&E scores.

Results

ID&E Scorecards

Component scores. Each ID&E observation yielded 13 individual component scores on a 4-point scale. The 13 component scores were calculated into a mean score for each teacher per observation. An overall mean score was calculated per teacher based on the two ID&E observations conducted. A campus mean score, standard deviation, and range were calculated using all teachers' individual overall ID&E mean scores. The results of the calculations are presented in Table 6.

Table 6

ID&E Scores and Teaching Performance Levels by Mathematics Teacher

Teacher	Grade	Tchr. Mean Score	T Level	Campus Mean	Campus Std. Dev.
AMT1	3	3.50	3	3.15	0.44
AMT2	4	3.19	3		
AMT3	3	3.85	4		
AMT4	5	3.54	3		
BMT1	3	2.81	1	3.19	0.29
BMT2	5	3.39	3		
BMT3	4	3.39	3		
BMT4	4	2.69	1		
BMT5	3	3.62	4		
BMT6	5	3.42	3		
BMT7	3	2.77	1		
BMT8	3	2.89	1		
CMT1	3	3.25	3	2.93	0.51
CMT2	4	2.92	2		
CMT3	3	1.96	1		
CMT4	4	2.04	1		
CMT5	5	2.89	2		
DMT1	4	2.89	2	2.99	0.38
DMT2	5	3.19	3		
DMT3	3	2.93	2		
DMT4	3	2.35	1		
EMT1	5	3.54	4	3.22	0.26
EMT2	4	3.50	4		
EMT3	3	3.62	4		
EMT4	3	2.85	1		
EMT5	4	3.43	3		
EMT6	5	3.31	3		
EMT7	3	3.15	2		

The four T levels were defined by how each mathematics teacher’s overall ID&E mean score compared to the campus mean and the standard deviation. Teacher mean scores more than one SD below the campus mean were considered T level 1; scores within one SD below the campus mean were considered T level 2; scores within one SD above the campus mean were considered T level 3; and scores more than one SD above the campus mean were considered T level 4. By converting all 28 teacher mean ID&E scores into T level scores, a better comparison could be made between the relative performances.

Across the five schools Table 6 and Figure 1 show that 12 of the 28 mathematics teachers were rated below their campus mean scores (T levels 1 or 2) and 16 were rated above their campus mean scores (T levels 3 or 4). All four of the mathematics teachers in Grades 3-5 at Campus A were rated above the campus mean. At Campus B four of the eight mathematics teachers in Grades 3-5 were rated below the campus mean (all ranked at T level 1) and four teachers were rated above the mean. At Campus C four of the five mathematics teachers were rated below the campus mean. At Campus D three of the four mathematics teachers were rated below the campus mean. At Campus E five of the seven mathematics teachers in Grades 3-5 were rated above the campus mean, with three of those ranked at T level 4. Each campus distribution of scores and T levels is shown in Figure 1 where the overall ID&E mean scores of the 28 mathematics teachers are indicated along with the T levels that are defined by the overall campus mean score and standard deviation.

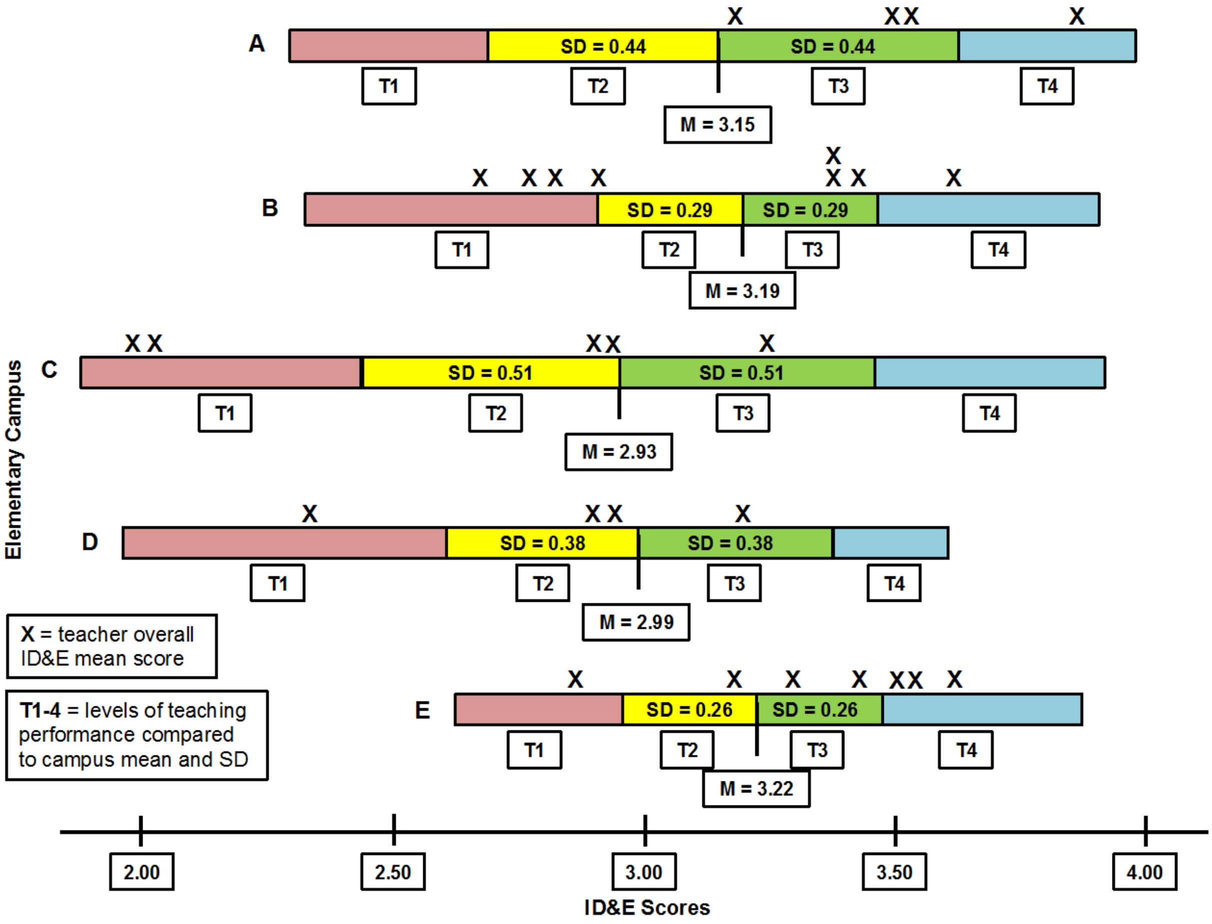

Figure 1. Mathematics teachers' ID&E overall mean scores and T levels by campus.

In Table 7, the mean score for each T level is reported per rubric component. There were eight teachers (16 ID&E scorecards) from four of the five campuses that made up T level 1. The T level 2 group was made up of five teachers (10 ID&E scorecards) from three of the five campuses. The T level 3 group was made up of ten teachers (20 ID&E scorecards) representing all five campuses. The T level 4 group was made up of five teachers (10 ID&E scorecards) from three of the five campuses. The information in Table 7 is represented graphically in Figure 2.

Table 7

Mean FFT Component Scores by T Level

FFT Domain/Rubric Component	T Level 1 Mean	T Level 2 Mean	T Level 3 Mean	T Level 4 Mean
1e	2.69	2.80	3.25	3.50
2a	2.44	2.90	3.60	3.70
2b	2.50	3.00	3.65	3.90
2c	2.69	3.10	3.45	3.80
2d	2.75	3.00	3.55	3.80
2e	2.56	3.00	3.15	3.30
3a	2.63	3.00	3.50	4.00
3b	2.25	2.50	3.30	3.70
3c	2.63	2.80	3.15	3.80
3d	2.38	2.80	3.35	3.60
3e	2.50	3.30	3.25	3.30
4a	2.75	3.00	3.20	3.40
4b	2.31	3.20	3.20	3.30

Note. Rubric components are labeled with the number of the domain and letter of the specific component, e.g. 3c = Domain 3, Component c ‘engaging students in learning’. All FFT components are presented in Table 1 (chapter 2).

The component mean scores for each T level reveal similar patterns. For example, in Domain 3 ‘instruction’ Component 3b ‘using questioning and discussion techniques’ was one of the lowest rated components at each level. The means scores for Component 3b stand in contrast to the mean scores for Component 3a ‘communicating with students’, which were higher at every T level. Components 2a ‘creating an environment of respect and rapport’ and 2b ‘establishing a culture for learning’ were among the highest rated components in Domain 2 ‘the classroom environment' at T levels 2-4. However, at T level 1 Components 2a and 2b had the lowest mean scores among the components of Domain 2.

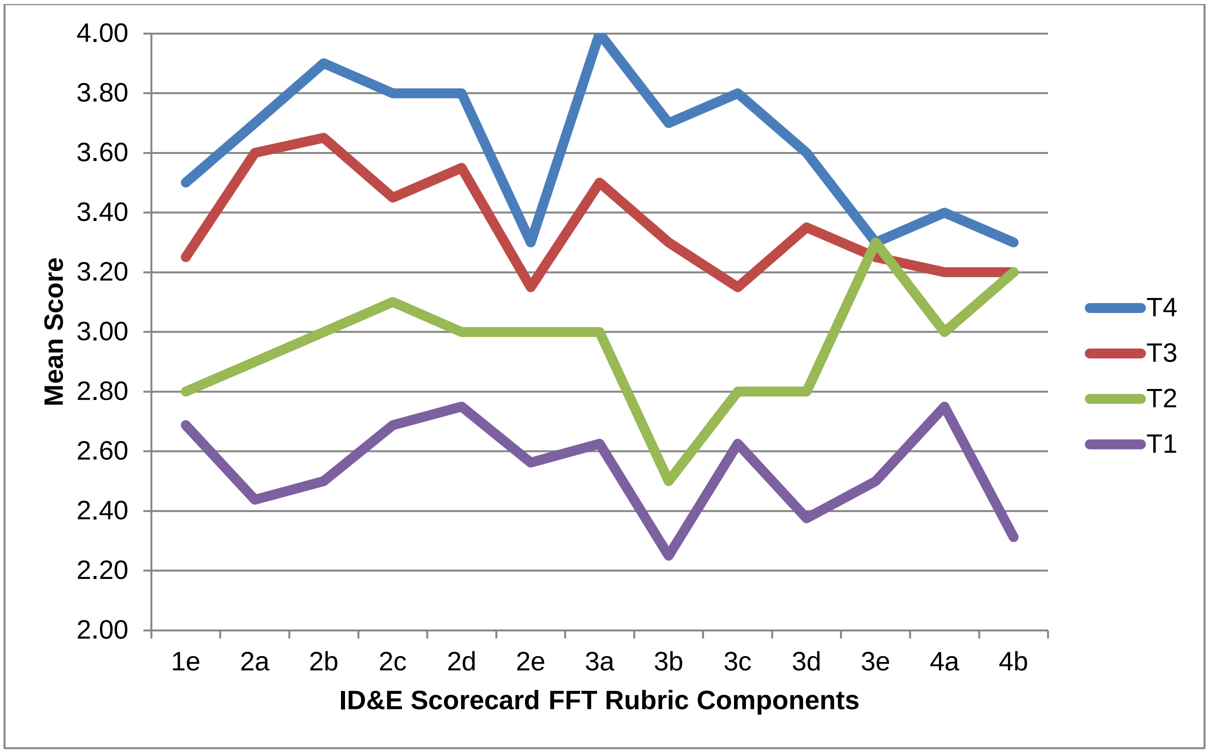

Figure 2. Mean FFT component scores by T level

Rubric evidence. A review of the 56 ID&E scorecards (both observations of the 28 mathematics teachers in Grades 3-5) revealed a wide range of evidence statements. Overall the volume of observation evidence for individual rubric components ranged from a single statement to over 20 statements. There appeared to be no relation between the number of evidence statements and the selected level of performance. The level of detail in the evidence statements also varied. On some scorecards numerous direct teacher and student quotes were cited along with very specific descriptions of the ongoing instruction, while on other scorecards evidence statements read like general summaries of what was heard and seen. The focus of the observer varied too; some observers focused entirely on the teacher's action and speech, while others also cited students’ actions and speech. As with the volume of evidence statements, both the level of details and the observer focus showed little relation to the levels of performance selected. In a very few cases, no evidence was cited for individual components, although levels of performance were selected.

Component 1e ‘designing coherent instruction’. The elements comprising this component, according to the FFT, include planning instructional groups, coordination of instructional materials and activities, and the alignment of a lesson within a curricular framework or unit. Evidence for this component was gathered during the pre conference. The majority of evidence cited for this component was summary in nature and addressed the alignment of the lesson within the curriculum and the appropriateness of the selected activities: “Learning activities are matched to instructional outcomes”; “All of the activities were the same which didn't allow for varied use of instructional groups or differentiation.” Neither scorecard quoted here went on to specify which particular activities were either matched to outcomes or were judged to be too similar.

Component 2a ‘creating an environment of respect and rapport’. The elements comprising this component include teacher interactions with students and student interactions with each other. Evidence for this and the other four components of Domain 2 was gathered during the classroom observation. Nearly all evidence focused on the degree of politeness and support shown in classroom interactions. Some of the evidence statements included: “T. is very positive and exhibits a caring demeanor towards her students”; “Student yelled at teacher ‘You called him’ and teacher redirected him with a look”; and, “Not all students respond with respect to the teacher, oftentimes multiple cues of redirection have to be given.”

Component 2b ‘establishing a culture for learning’. The elements comprising this component include how the teacher frames the importance of the content, the expectations for achievement, and students’ pride in their work. The two most common themes in the evidence statements for this component were how teachers explained the everyday application of the mathematical concept and the expectations for what counts as excellent work in the classroom. Statements included: “The teacher reminded the students why it is important to take notes”; “Teacher is committed to learning, but has difficulty getting students to be committed to learning as well”; “T. related topic/concept to their personal lives”; “T: ‘If you go to the store and want to buy something, you will need to be able to quickly round estimate to see if you have enough money”; and, “T. constantly reminded the students that with hard work, they can be successful.”

Component 2c ‘managing classroom procedures’. The elements comprising this component include the management of instructional groups, materials, routines, and transitions between activities. The most common theme that emerged in the evidence for this component was how classroom routines affected instructional time. Statements included: “Restroom procedure is in place and does not stop instruction”; “Students began with a fact table completion in an allotted 10 min. time frame. Students call out done and teacher tells the student the time”; and, “Smooth transitions between activities ensure maximum instructional time.”

Component 2d ‘managing student behavior’. The elements comprising this component include the classroom expectations for behavior, management of behavior, and response to misbehavior. The most common theme that emerged in the evidence statement for this component was how teachers responded to misbehavior and redirected students. Statements included: “Teacher call student name and takes item from student. Then he says thank you”; “Teacher attempts to maintain order in the classroom but has uneven success”; “The teacher asked the student to please stop yelling out”; and, “No true misbehavior was observed.”

Component 2e ‘organizing physical space’. The elements comprising this component include the arrangement of furniture in the classroom, use of resources including technology, and classroom safety. While the arrangement of furniture and the general safety of the classroom are generally static features and they were noted in the evidence, most observations focused on the use of technology during instruction. Evidence statements included: “The teacher utilizes the promethean board and available online resources”; “Promethean board was used by both teachers and students”; and, “Brain Pop was used to reinforce the concept.”

Component 3a ‘communicating with students’. The elements comprising this component include teachers’ use of oral and written language, explanations of content, directions for activities, and expectations for learning. Evidence for this and the other four components of Domain 3 was gathered during the classroom observation. This component typically yielded the most evidence including numerous direct quotes from teachers. Evidence statements for this component were frequently double coded and repeated as evidence for other components such as 3b. The two most common themes that emerged in the evidence for this component were teacher explanations of content and teacher explanations of instructional procedures. Statements included:

“’Remember, you guys should be talking to each other’”; “Teacher summarizes main points and clarifies students’ responses”; “Teacher uses concepts and language the students understand”; “Teacher explained what they needed to do on the hundreds chart”; “Teacher explained why it’s important to know and how we use elapsed time in everyday situations”; and, “’So as you can see, there are many ways to get the answer’.”

Component 3b ‘using questioning and discussion techniques’. The elements comprising this component include the quality of questions and prompts, discussion techniques, and student participation. Evidence statements for this component focused primarily on actual questions that teachers asked to facilitate understanding. Another common theme that emerged in the evidence was how students were invited or prompted to answer a question or participate in a discussion. Statements included: “Students called consecutively to answer the question or hold the pieces up to show a reflection or not a reflection”; “The teacher drew names to call on the students”; “Stop and think/turn and talk was used”; “T: ‘Tell us why you are boxing those problems’”; “Can you think of any place you would see an array?” and, “Where did the five come from?”

Component 3c ‘engaging students in learning’. The elements comprising this component include the structure and pacing of a lesson, the instructional materials and resources, the grouping of students, and the activities and assignments. The most common theme that emerged in the evidence for this component was the degree to which students were on task and participating in the lesson. Statements included: “During the discussion 4 of the 5 groups were discussing the patterns, while the other group was not engaged in a discussion"; "Some students are intellectually engaged in the lesson"; "23/24 students were actively engaged"; and, "Most students enthusiastically participate."

Component 3d 'using assessment in instruction'. The elements comprising this component include assessment criteria, teacher monitoring of student learning, teacher feedback, and student self-assessment and monitoring of progress. The most common themes that emerged in the evidence for this component were how teachers gauged student understanding and how they provided feedback to students. Statements included: "Thumb check was used to see if students understand"; "Students were instructed to get out their journals and write today's learning"; "Teacher worked with small group of students to reinforce and review the concept"; "Teacher observes a student incorrectly modeling, works with that student independently, circulates to work with other students, then goes back to the student who did it incorrectly and asked him to model again"; and, "T: 'Show me how you measured it'."

Component 3e 'demonstrating flexibility and responsiveness'. The elements comprising this component include lesson adjustment, teacher persistence, and teacher response to students. The most common theme that emerged in the evidence for this component was how teachers re-taught a concept when student confusion or misunderstandings were encountered. Statements included: "T: 'So, to make it easier, label the ruler'"; "The teacher, realizing the students needed additional support, provided additional strategies"; "Teacher makes an effort to help struggling students and has additional strategies and tools they can use"; "Teacher retaught the concept that students did not quite understand"; and, "T: 'I want to see how you're thinking through this'."

Component 4a 'reflecting on teaching'. The elements comprising this component include the accuracy of teacher reflections and explanations of use in future teaching. Evidence for this component was gathered during the post conference. Evidence statements were generally summary in nature, most frequently mentioning that teachers reflected on their instruction, with few details on what exactly they said during the reflection. When details were presented, they tended to focus on how lessons could be better paced and how students could be more involved. Several scorecards showed no evidence for this component. Statements included: "Teacher reflects on her lessons daily and has ideas about how they can be improved"; "The teacher recognized that students should have more accountable talk and interactive learning within the groups"; and, "T: 'I would plan to make the lesson last two days. I would have them round to the nearest half one day and round to the nearest quarter the next day."

Component 4b 'maintaining accurate records'. The elements comprising this component include how teachers maintain records of student completion of assignments and progress and how they maintain non-instructional records. Evidence for this component was gathered during the classroom observation, during the post conference, and throughout the year. The most common themes that emerged in the evidence for this component were teachers maintaining their lesson plans and grade books, and how assessment data are maintained. Statements included: "Teacher and students maintain records citing master and non mastery"; "Teacher consistently maintains her grade book and lesson plans"; "Teacher consistently maintains her lesson plans"; "T: 'I keep a binder with student information. Each student also has a binder and they log in their work and their grade"; "Teacher has a data binder for each student and uses this information to track student progress"; and, "Teacher has a data binder, and has an effective system in place to monitor student progress."

In summary, written entries across components indicate general agreement among administrators on what counted as relevant evidence. However, levels of detail and specificity in the evidence statements varied.

Administrator Interviews

The analysis of the interview responses centered on the topics identified in the five questions from the interviews.

Elements of effective mathematics instruction. The first interview question that administrators were asked was, "What do you typically find in a mathematics classroom where the lesson seems to be working well?" This introductory question asked administrators to reflect overall on mathematics instruction at all elementary grade levels and to describe effective practices without necessarily using the language of the FFT or referring to the rubric components.

One of the most common responses to this question, mentioned by seven of the ten administrators, was the use of small groups where students can work with each other, talk to each other, and collaborate on solving problems. As one administrator stated, "I typically find students who are in small groups collaboratively working with one another, and they're conversing with each other, and they have manipulatives, and they are exploring with one another." Another common response, also mentioned by seven administrators, was the use of manipulatives in the classroom so that students could have hands-on experience representing and solving problems: “Hands-on activities and the use of manipulatives, the teacher is modeling and the students have a chance to talk and work it out, well first to understand what the problem is asking them, and then work out the problem using the manipulatives and then transfer that information to their papers.” Other responses included teachers being fully planned and prepared, teachers adjusting instruction when misconceptions or misunderstandings are encountered, the use of formative assessments, and the use of journals.

Most effective observed lessons. The second interview question was, “In the most effective mathematics lessons you observed, which ID&E rubric components did you rate the highest and why?” Administrators were free to mention specific teachers by name and cite evidence from the ID&E scorecards. About half of the administrators looked back on their notes and scores; the others spoke based on what they remembered about the highest-rated lessons.

The three most common responses to this question, mentioned by at least five of the administrators, did not focus directly on teacher instructional techniques. One theme was lessons were consistently well planned and prepared with a particular focus on differentiating instruction (FFT Component 1e ‘designing coherent instruction’): “He has three rotations and he knows I’ve got a high group, I’ve got a medium group, and I’ve got a low group, and to see how he takes the same lesson and differentiates it among his rotation.” Another theme that emerged was teachers had established a positive rapport with students (Component 2a ‘creating an environment of respect and rapport’): “She has a great rapport with the students. Once the kids understand that the teacher cares for them, they will be willing to meet those expectations.” The third theme was teachers had set the expectation that students would be able to articulate their mathematical thinking in classroom discussions to help each other learn (Component 2b 'establishing a culture for learning'): "She has a good relationship with the kids where they're very, um – they come up to the board, if somebody's stuck they call on each other to help each other, she really has taught them to utilize each other as support versus her giving them all the support and her giving them all the answers, so they used each other a lot of the time to guide each other through the lesson." Another administrator described it this way: "He really questions the children a lot and forces them to articulate what they're thinking. That's difficult, especially if they come with deficits, so the more they talk, the more they understand." Other responses to this question cited Component 3d 'monitoring student progress', Component 3e 'demonstrating flexibility and responsiveness' (i.e. lesson adjustment), Component 2c 'managing classroom procedures', and Component 2d 'managing student behavior'.

Least effective observed lessons. The third interview question was, "In the least effective mathematics lessons you observed, which ID&E rubric components did you rate the lowest and why?" Again, administrators were free to mention specific teachers by name and cite evidence from the ID&E scorecards. None of the administrators looked back on their notes and scores; they all spoke based on what they remembered about the lowest-rated lessons.

There was very little agreement among the ten administrators on what constituted an ineffective lesson. It seemed that any FFT component, if severely lacking or under-developed, was sufficient to rank the lesson as ineffective. Three of the administrators cited a lack of planning (Component 1e): "That is actually kind of easy. I think that the one where they're not doing well is actually the designing instruction – coherent instruction." Three administrators also cited classroom discussions, or the lack of them: "For Domain 3, 3b using questioning and discussion techniques, a lot of times there was discussion, but it was really facilitated by the teacher and the teacher would ask one student a question or they would address the whole group and one student would answer."

The rest of the responses to this question were evenly split between components in FFT Domain 2 'the classroom environment' and components in Domain 3 'instruction'. Domain 2 components mentioned included 2c 'managing classroom procedures', 2a 'creating an environment of respect and rapport', and 2b 'establishing a culture for learning.' Responses included: "That's going to go ahead and affect Components 2a and 2b because if I'm not respecting the learner and if I'm not establishing that culture for learning because I'm not prepared, they're not going to learn to love this content or this subject area"; "A lot of time was spent on just low level repetition and, you know, I see that as a loss of instructional time, not moving forward when you see that the kids are ready"; "2b has been an area that is very specific to weak teachers, the culture for learning. Again I say that's how you set up your class – it's over time, it's not a one-time deal"; and, "With 2a – creating an environment of respect and rapport, there's no sense of cohesiveness or family. There's no respect. The students will be just unhappy and not interested in the content, it doesn't matter what it is."

Other Domain 3 components mentioned as responses to this question, besides 3b, included 3d 'using assessment in instruction', and 3a 'communicating with students'.

Responses included: “I would also include 3d – assessment – your assessment techniques, you know, using formative assessment throughout the lesson to gauge the understanding and the learning so you’ll know – that’s something concrete to let you know – whether or not you need to make adjustments or modifications”; and, “Oh, and his feedback also: ‘very good’ but not really telling them why, ‘good job’ and keeping right on, but we need constructive feedback where they’re really understanding why they didn’t get it right or what they’re doing wrong.”

Most essential FFT components. The fourth interview question was, “Which rubric components do you think best capture effective mathematics instruction?” All of the administrators referred to the completed ID&E scorecards to help them formulate answers. The most common component cited in response to this question, mentioned by six administrators, was 3c ‘engaging students in learning’: “If the kids are engaged in something and it’s meaningful, then everything else will fall into place. I mean, maybe I’m minimizing this – it’s really complex actually with how everything is so interwoven, so I know the kids know when they’re engaged.” Another administrator said, “If students aren’t engaged in the activity, so they have to have something that’s meaningful to them to help them make the connection, and a lot of times if we – if they’re just given busy work, then that’s not actual student engagement.” Several administrators emphasized the point that student engagement is not accidental; it’s the result of good planning.

The other two most common components cited in responses to this question, each mentioned by five administrators, as being hallmarks of effective mathematics instruction were 3a ‘communicating with students’ and 3b ‘using questioning and discussion techniques’. Responses included: “I think 3a, communicating with students, because if the students don’t understand what you’re saying or if they can’t follow how to process, or if they’re not asked the right questions or shown different ways of looking at things, it kind of limits them and it prevents them from even being able to go to the questioning and discussion techniques”; and, “I would also say 3b, which this is, as a campus, this is one of the areas that we are struggling with – the questioning part – because they have to be able to ask the appropriate questions to lead the students into good discussion in order for them to make connections and discover different things.” One administrator made two points emphasizing the importance of Component 3b: “In order for us to improve with mathematics, we gotta get that discussion going, we gotta have to have these children talking mathematics, using math terms, and it being just a regular conversation, a normal conversation”; and, “A teacher that can use questioning and discussion techniques to get the kids thinking about math is going to be that teacher that is really, really making the difference with how our children are obtaining the math skills that they need.”

Other responses to this question cited components 2a ‘creating an environment of respect and rapport’ and 3d ‘using assessment in instruction’. Specific responses and reasons included: “Once the kids understand you care for them, they’ll do anything for you. So, that is crucial. If that’s not there, it’s going to cause problems with any lesson”; “That would be one where you’re creating that environment where they respect each other, they respect you, they respect your authority, and they’re willing to listen to the actual lessons”; “If you’re not formative assessing them, using formative assessment throughout the lesson, then at the end you if have a summative assessment then that was really kind of a waste of a lesson.”

Greatest campus need. The fifth interview question was, “Which rubric component would you identify as an area of general need for your math teachers?” None of the administrators referred to the completed ID&E scorecards in answering this question. The three most common responses, each cited by at least half of the administrators, were Components 1e ‘designing coherent instruction’, 3b ‘using questioning and discussion techniques’, and 3c ‘engaging students in learning’. A theme that emerged is how inter-related the FFT components are and how concentrated improvement in one component can be reflected in other components. One administrator explained how components 3b and 1e are related: “And with the questioning one, we’ve also been able to hit some of the designing coherent instruction because we talked about at our PD, if you don’t plan for those questions, and you don’t have them ready and you don’t really look at what you’re going to question the students on during the lesson, then it’s not going to be a successful.” Another administrator explained how Components 3b and 3c are related: “Those were the areas that were typically scored the lowest because they’re so closely related. For me, that discussion and that questioning really is that engaging piece because those are the things that allow the students the opportunity to think about what they’re doing.”

VAM Data

The VAM analysis data are presented in Table 8. For each grade-level mathematics cohort of students (15 cohorts in total across the five campuses), the NCE gain was reported along with the standard error for the calculations and the index, which was calculated by dividing the NCE gain by the standard error, were reported to the district. Similar to the teacher ID&E overall mean scores, the VAM scores were converted into ordinal data. As described in Table 4 (found in chapter 3), NCE gains that were more than one SE below zero (with an index score of -1.0 and below) were considered V level 1, or ‘less than expected growth’; NCE gains that were between one SE above and below zero (with an index score between 1.0 and -1.0) were considered V level 2, or ‘expected growth’; and NCE gains that were more than one SE above zero (with an index score of 1.0 and above) were considered V level 3, or ‘greater than expected growth’. Eight grade level cohorts scored at the V1 level, three cohorts scored at the V2 level, and 4 cohorts scored at the V3 level. As mentioned previously, due to district policy, teacher-level VAM scores were not reported. As a result, every teacher at a particular grade level was assigned the cohort score and V level.

Table 8

Value-Added Scores and Levels for Mathematics by Campus and Grade

Campus	Grade	NCE Gain	Std. Error	Index	V Level
A	3	-8.19	1.80	-4.55	1
A	4	-10.02	1.86	-5.37	1
A	5	2.01	1.68	1.20	3
B	3	-2.42	1.52	-1.59	1
B	4	-2.07	1.43	-1.45	1
B	5	3.61	1.23	2.95	3
C	3	-0.11	1.69	-0.06	2
C	4	0.67	1.35	0.49	2
C	5	-6.19	1.61	-3.83	1
D	3	-15.60	1.94	-8.03	1
D	4	-10.56	1.72	-6.14	1
D	5	1.83	1.86	0.99	2
E	3	-5.64	1.51	-3.72	1
E	4	6.44	1.62	3.98	3
E	5	3.15	1.75	1.80	3

Figure 3 shows this information for the third grade cohorts at the five elementary schools studied. The V levels are delineated as determined by the standard error from the VAM calculation (the distance of the SE both above and below 0), and the cohort's mean NCE gain is noted. Figures 4 and 5 show this information for the fourth grade and fifth grade cohorts respectively.

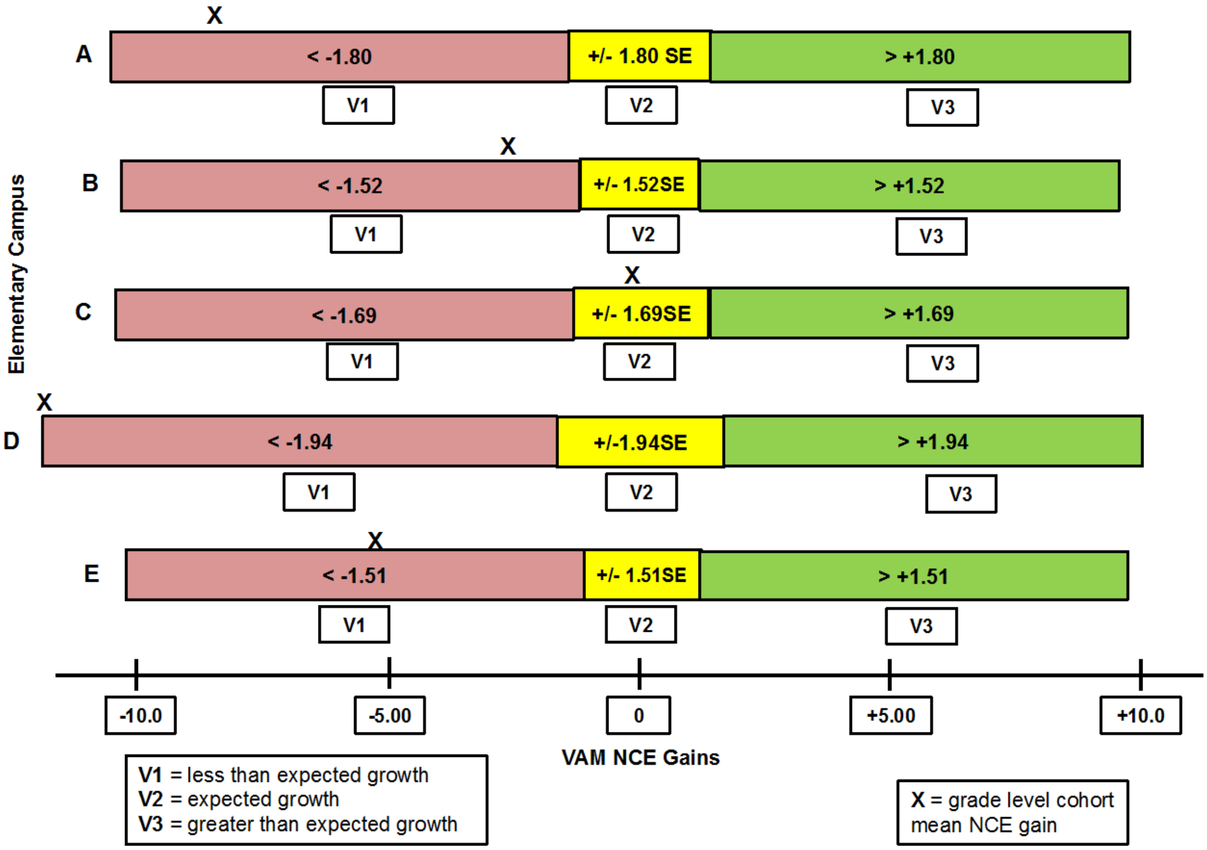

Figure 3. Value-added scores and levels for 3rd grade mathematics cohorts.

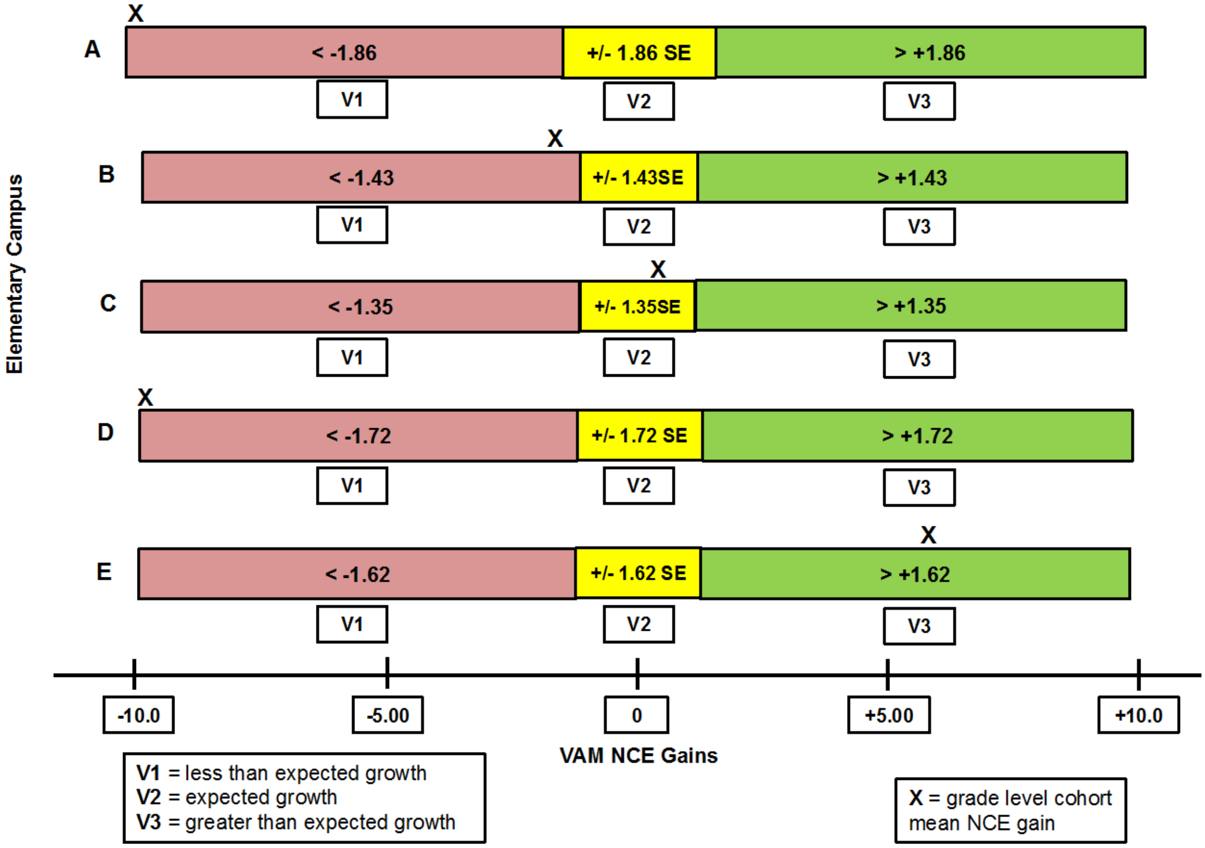

Figure 4. Value-added scores and levels for 4th grade mathematics cohorts.

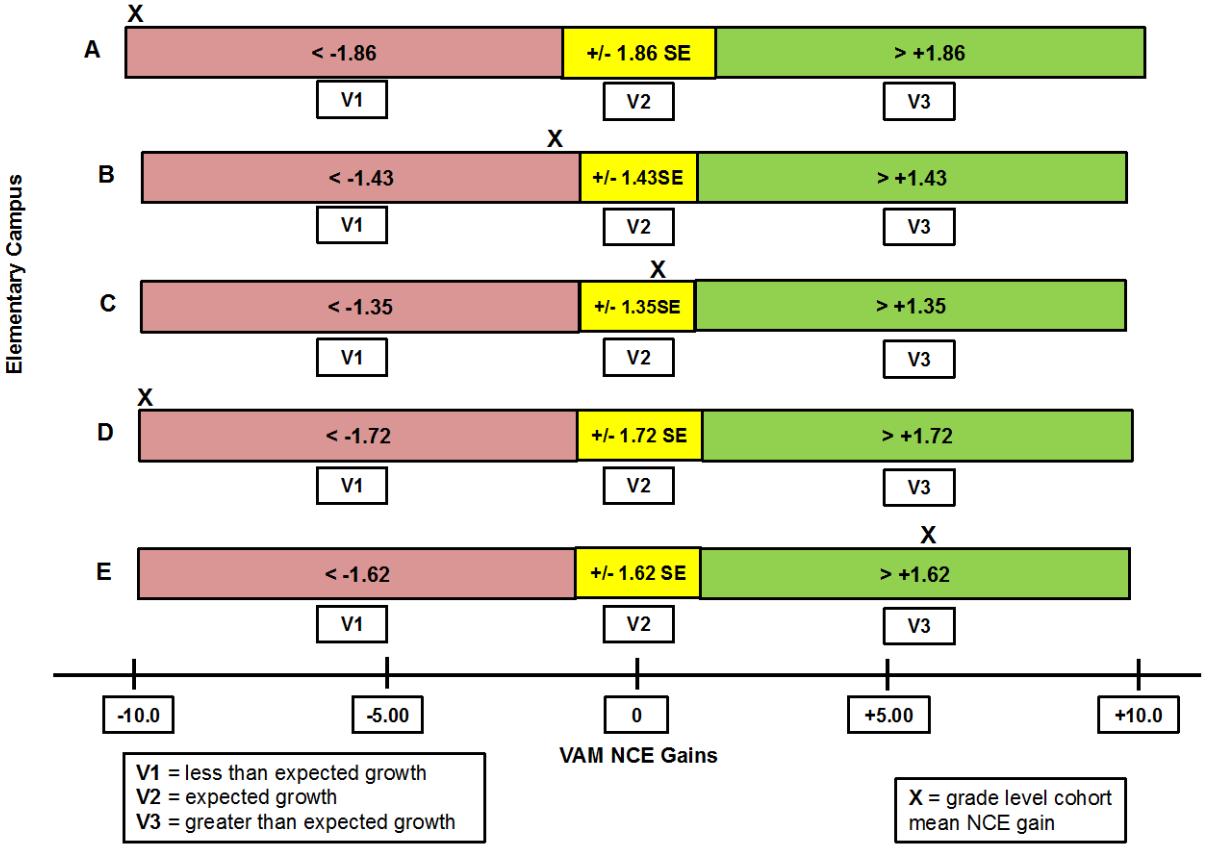

Figure 5. Value-added scores and levels for 5th grade mathematics cohorts.

Answers to Research Questions

1. What is the relationship between students' progress scores in mathematics, as measured by VAM, and their teachers' instructional performance scores, as measured by the FFT?

A correlation test was run in SPSS using the 28 T and V levels. The results are summarized in Table 9.

Table 9

Correlations of Teaching Performance Levels and Value-Added Levels

			T Level	V Level
Spearman's rho	T Level	Correlation Coefficient	1.000	.350
		Sig. (2-tailed)	.	.068
		N	28	28
	V Level	Correlation Coefficient	.350	1.000
		Sig. (2-tailed)	.068	.
		N	28	28

The correlation of .350 was modest at best, and it was not a statistically significant relationship with a p-level of .068. The relationship between teachers' ID&E overall mean scores based on the FFT ratings (T levels) and their students' progress, measured by VAM (V levels), was not established here. While there were instances of teachers with both low T and V level rankings and teachers with both high T and V level rankings, most of the teachers had T and V level rankings that did not match, as seen in Figure 6. This illustrates the relatively weak relationship shown between the two measures in this study.

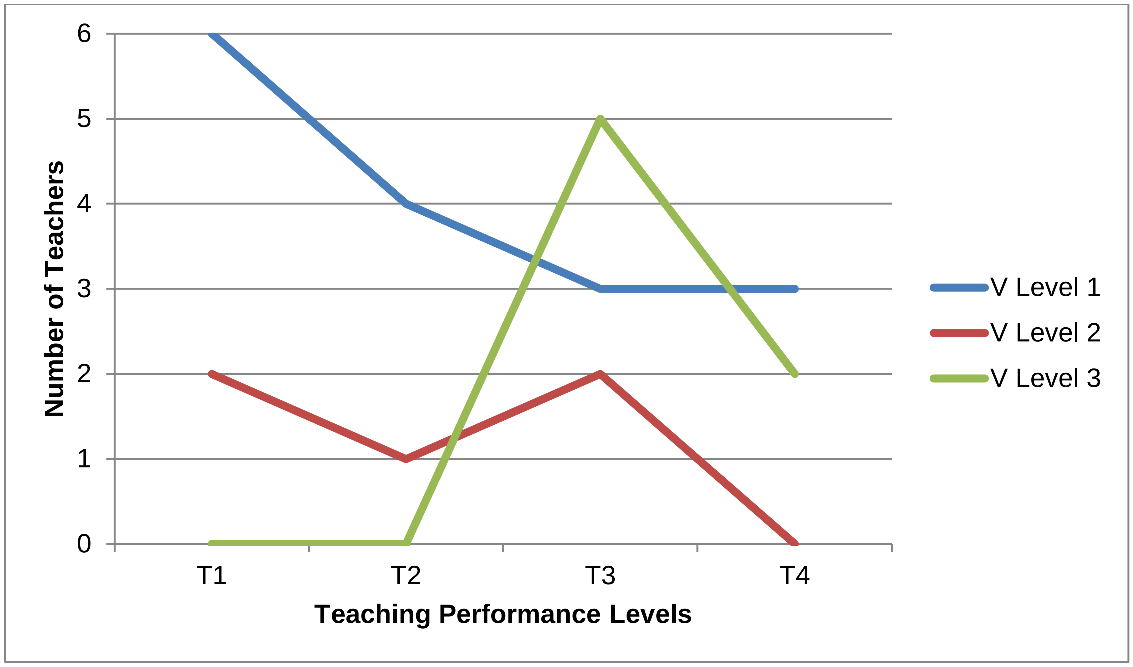

Figure 6. Correlations of teaching performance levels and value-added levels

2. What is the relationship between levels of teacher and student performance, as articulated in interviews with campus administrators, and an examination of their documented observation evidence using the FFT?

The matrix of performance measures first outlined in Table 5 (chapter 3) is presented in Table 10 with the number of paired samples into each cell of the combined metrics. As mentioned above, seven teachers were rated high V/T with a V level of 3 and T level of either 3 or 4. No teacher who earned a V level of 3 also had a T level of 1 or 2, which would have been a mismatch of measures. A mismatch occurred elsewhere in the matrix, however, where three teachers were rated at the highest T level (4) and the lowest V level (1). In all three cases, the teacher was part of a team of third grade teachers; sizes of the teams ranged from two to four teachers.

Table 10

Distribution of Mathematics Teachers in the Matrix of Performance Measures

	Level V1 NCE Gain More than 1 SE below 0	Level V2 NCE Gain Between 1 SE above/below 0	Level V3 NCE Gain More than 1 SE above 0
Level T4 Teacher ID&E Score More than 1 SD above campus mean	*High T/Low V* T4V1 *n* = 3	*High T/Med V* T4V2 *n* = 0	*High T/High V* T4V3 *n* = 2
Level T3 Teacher ID&E Score Within 1 SD above campus mean	T3V1 *n* = 3	T3V2 *n* = 2	T3V3 *n* = 5
Level T2 Teacher ID&E Score Within 1 SD below campus mean	*Low T/Low V* T2V1 *n* = 4	*Low T/Med V* T2V2 *n* = 1	*Low T/High V* T2V3 *n* = 0
Level T1 Teacher ID&E Score More than 1 SD below campus mean	T1V1 *n* = 6	T1V2 *n* = 2	T1V3 *n* = 0

Ten of the 28 teachers were rated as low T/low V with a T level of either 1 or 2 and a V level of 1. Seven teachers were rated as high T/high V with a T level of either 3 or 4 and a V level of 3. The five teachers in the middle column with a V level of 2 were split with regard to their T levels – two were rated T3, indicating high T/med V, and the other three were rated as either T1 or T2, indicating low T/med V. There were six teachers with mismatched levels rated as high T/low V with a T level of either 3 or 4 and a V level of 1. No teachers were rated as low T/high V with a T level of either 1 or 2 and a V level of 3.

Table 11 shows the overall mean ID&E ratings that the ten administrators made overall for all of their campus teachers along with the mean ID&E ratings they made for their mathematics teachers in Grades 3-5. The administrators' years of experience using the FFT is also reported along with indications of whether they taught mathematics in Grades 3-5 or served as an elementary mathematics coach.

Table 11

Administrator Observation Mean Scores and Experience

Administrator	Overall Mean Ratings	3-5 MT Mean Ratings	Yrs. FFT Experience	3-5 MT Experience	Elem. MC Experience
AAD1	3.26	3.73	1	N	N
AAD2	3.03	3.31	1	N	Y
BAD1	3.20	3.13	3	N	N
BAD2	3.18	3.12	3	Y	Y
CAD1	2.86	2.54	3	N	N
CAD2	3.00	3.31	2	N	N
DAD1	2.94	2.83	2	Y	N
DAD2	3.04	2.85	1	Y	N
EAD1	3.30	3.38	3	Y	N
EAD2	3.15	3.31	3	Y	Y

The administrators at Campus B and Campus E all had three years' experience using the FFT and had mean ratings on both measures that were within 0.15 of their campus colleague. Additionally, one of the administrators at each campus previously served as an elementary mathematics coach. The administrators at Campus D had mean ratings on both measures that were similarly close to each other, within 0.15. While they had only two and one years experience using the FFT, each had previously taught mathematics in Grades 3-5. At Campus A, neither administrator had experience teaching mathematics in Grades 3-5 – one taught mathematics in the primary elementary grades and the other taught mathematics at the middle school level (and also served as an elementary mathematics coach). Their mean ratings were 0.23 apart overall and 0.42 apart for their mathematics teachers in Grades 3-5. At Campus C neither administrator taught mathematics in Grades 3-5 or served as an elementary mathematics coach. Their mean ratings for their mathematics teachers in Grades 3-5 were 0.77 apart and 0.14 apart overall.

Chapter Summary

In this chapter, the data from the various sources were reported, analyzed, and considered in relation to the research questions. In the next chapter, the study findings are interpreted in relation to the main research questions. Other relationships that developed among the data and the limitations and implications of this study are also discussed. Chapter 5 concludes with recommendations for future researchers investigating the relationship between multiple measures of effective teaching.

CHAPTER 5

DISCUSSION

The purpose of this study was to investigate the extent to which teacher performance and student performance measures correlated, and to understand which specific practices of mathematics teachers related to student performance. In this chapter, the findings are summarized and interpreted in relation to the main research questions. Other relationships, limitations, and implications of this study are also discussed. This chapter concludes with recommendations for future research.

The following two research questions guided the study:

1. What is the relationship between students' progress scores in mathematics, as measured by VAM, and their teachers' instructional performance scores, as measured by the FFT?
2. What is the relationship between levels of teacher and student performance as articulated in interviews with campus administrators, and an examination of their documented observation evidence using the FFT?

The hypotheses/assumptions were as follows:

1. It was hypothesized that there would be a significant relationship at the $p < .05$ level between teacher performance FFT scores and student performance VAM scores.
2. It was assumed that students' value-added progress measures would positively relate to teacher instructional performance ratings. Specifically, it was assumed that teachers rated high according to their FFT observation scores would also be rated high according to their VAM student progress scores, and teachers rated

low on FFT would also be rated low on VAM. It was also assumed that interviews with school administrators and an examination of their observation documentation would provide insight into confirmatory or discrepant VAM/FFT relationships.

The findings revealed that the hypothesis in the first question was not supported, rendering a moderate relationship between teacher FFT ratings and students' VAM scores. However, the findings for the second question, gained in analyses of raters' teaching experience, years in the field of administrative leadership, interview responses, and cited observation evidence, helped to explain both confirmatory and discrepant VAM/FFT relationships.

Interpretation of the Findings

One of the factors that possibly affected the relationship between the FFT and VAM measures was the lack of teacher-level VAM data available from the district. While the T level measure was earned individually, the V level measure was earned collectively, in most cases. The categorization of individual teachers as effective based on comparing an individual measure to a collective measure renders the comparison suspect. While all teachers who earned the highest V level also earned an above average T level, not all teachers who earned an above average T level also earned the highest V level. In fact, three teachers who earned the highest T level individually were part of teaching teams that earned the lowest V level. One of those three teachers, BMT5, had three grade level colleagues who also contributed to the collective V level 1 score, and all three of those colleagues were rated at T level 1. Teacher EMT3, who was rated at T level 4, had two grade level colleagues, both rated below average – one at T level 2 and one at T level 1. Curiously, in the third case, teacher AMT3, who was rated at T level 4, had only one grade level colleague and that teacher was rated above average at T level 3.

It could be the case that teachers BMT5 and EMT3 contributed equally to the collective V level 1 that indicated less than expected growth across their grade level cohorts of students, but it could also be argued that perhaps they didn't, given the wide difference between their individual T levels and those of their teammates. For example, an examination of the ID&E scorecards shows that teacher BMT5 scored at the distinguished level of performance for Component 3b 'using questioning and discussion techniques', the component frequently cited by administrators as being essential to effective mathematics instruction and the one shown in the overall data as being an opportunity across the campuses and all four T levels. Comments included: "Teacher uses open ended guiding questions to support student learning"; "Teacher uses wait time, allowing students time to articulate their understanding"; and, "Teacher calls on many students, even those who don't volunteer." By contrast, the three grade level colleagues of BMT5 scored at the basic performance level for Component 3b. Fewer evidence statements were listed on the scorecards and they included: "Limited discussion went on amongst groups about possible answers"; "Teacher framed some questions designed to promote student thinking, but most students did not go into a discussion"; and, "Questions are simple questions, and only a small number of students had discussions amongst each other." Only teacher level VAM data would be able to confirm whether or not the four teachers contributed equally to the cohort's V1 ranking, however, it seems likely that they did not.

Another interesting VAM result that may have contributed to the discrepancy between T and V levels was the distribution of V level scores across grade levels. All five third grade cohorts of students showed growth at one standard error below the mean; hence all third grade teachers were rated at V level 1. An explanation stems from the fact that third grade is the first grade level when Texas students take the STAAR test. VAM growth for third grade students is measured by comparing students' performances on the third grade STAAR test to their performances on the second grade Stanford 10 test, as previously mentioned in chapter 3. While the translation of scores from both tests into NCEs allows for a comparison of the two, the Stanford 10 is approached with a different purpose, and implications for students and schools differ from the STAAR.

In contrast, three of the four cohorts that showed greater than expected growth were at the fifth grade level where it is noted that teachers were rated at V level 3. Fifth grade is the third consecutive year that students take the STAAR test, so it's assumed that they have acclimated to it by then. It's possible that fifth grade is a 'bounce back' year where student cohorts recover some of their mean NCE gains that were lost in third and fourth grade during the transition to STAAR testing. Another explanation for the amount of VAM gains shown at the fifth grade level could be explained by teacher assignments, influenced by state policies. Fifth grade has long been the first grade where Texas students were required to pass state tests in reading and mathematics before matriculating to sixth grade. A common staffing practice at many elementary campuses has been the placement of strongest teachers in fifth grade to cope with those requirements.

Of the seven teachers rated high T/high V, only AMT4 was the sole mathematics teacher in Grade 5 at Campus A. The other six comprised three pairs of grade level colleagues: BMT2 and BMT6 teamed in Grade 5 at Campus B, EMT2 and EMT5 teamed in Grade 4 at Campus E; and, EMT1 and EMT6 teamed in Grade 5 at Campus E. While every teacher who was ranked as high V (V level 3) was also ranked as high T (T level 3 or 4), none of the high V teachers teamed with a colleague ranked as low T (T level 1 or 2). This could be evidence of intentional staffing by principals pairing up consistently strong teachers or it could suggest the benefit of other campus practices beyond the scope of this study such as collaborative planning. More in depth study of hiring practices would increase the understanding of this phenomenon.

Contextual Factors and Alternative Explanations

Administrator Experience

As noted in Table 2 (chapter 3), the ten participating administrators had varied levels of experience in school administration, teaching mathematics, and using the FFT as an observation instrument. These variations possibly affected what specifically was noted during observations (mathematics experience), how it was expressed in rubric evidence statements (FFT experience) and how feedback may have been delivered in post-conferences (administrative experience). While post conferences were not observed, based on ID&E scorecards, differences were noted in the volume, relevance, and level of detail of evidence statements. Administrators with experience teaching mathematics in Grades 3-5 generally provided more thorough evidence on scorecards. That evidence was enhanced based on greater familiarity with the FFT, more experience as an administrator, and, especially, when administrators had previously coached elementary mathematics.

All three previous elementary mathematics coaches scored the mathematics teachers in Grades 3-5 at their schools lower than their campus administrative colleagues. At campuses B and E, those previous coaches (BAD2 and EAD2) both had three years experience using the FFT and were paired with administrators who also had used the FFT for three years. The gap in rating averages (both for the mathematics teachers in Grades 3-5 and for all campus teachers) between EAD2 and EAD1 was small, while the gap between BAD2 and BAD1 was negligible (see Table 11, chapter 4). This suggests that prior experience in teaching the subject area coupled with experience using the observation rubric leads to more calibration among teamed observers. Further, half of the six grade level cohorts from those two schools scored at V level 3, possibly indicating that consistent, convergent, detailed feedback on teaching might positively impact instructional improvement and student growth.

The two most divergent scoring teams were at Campuses A and C, where none of the four administrators had previous experience teaching mathematics in Grades 3-5. Years using the FFT appeared not to resolve those differences as CAD1 and CAD2 had three and two years respectively, and AAD1 and AAD2 were both using the rubric for the first time. While AAD2 was a former mathematics coach and had worked with instruction in Grades 3-5, AAD1 was the only administrator both new to the rubric with no previous instructional experience or support in mathematics in Grades 3-5. Perhaps not unexpectedly, AAD1 had the highest mean ratings for mathematics teachers by far. It’s fair to assume that some of those ratings might have been overinflated due to lack of experience with content-specific instruction at those grade levels. The observer team at Campus D further supported this reasoning where, despite less than three years experience each using the FFT, both DAD1 and DAD2 were closely calibrated, and both had previously taught mathematics in Grades 3-5.

FFT as a Factor

As mentioned in chapter 2, the FFT is both celebrated and criticized by educators for its universal applicability. While it provides a common framework for observations and a common language to discuss instruction across grade levels and subjects, it does not provide content-specific lenses to examine details. For example, Component 3b ‘using questioning and discussion techniques’ is a critical component to instruction, as the ten administrators also noted. However, what counts as an effective set of questions in mathematics is very specific to mathematics learning and skill acquisition. That questions were asked in a lesson and noted in an observation does not reveal whether the questions were appropriately challenging, fostered thinking, or contributed to mathematics understanding. The three administrators who had previously served as mathematics coaches were likely at an advantage in recognizing when effective questions were effective mathematics questions, having repeatedly coached others how to formulate them. In their interviews, the former coaches each described questioning and its importance for building conceptual understanding in much greater detail than their colleagues. One emphasized the importance of mathematical processes and discussing how problems are solved: “I typically find the teacher using multiple strategies or ways to address an idea and allowing the kids to share their approach – either the kids sharing their approach and the teachers taking their ideas and kinda bringing them together to make sense for everybody, but just different approaches and conversations about how to approach something." Another noted from an observation: "As far as questioning, so it's not just 'What's the answer?', it's 'How did you figure that out?' 'Where did the 5 come from?' 'Can you give me an example?' 'Why do we use division?' 'Why would you need to divide?' Once again, she allows numerous opportunities for accountable talk, and then, of course, accountable talk was heard."

Demands on Administrators

Each of the five elementary schools in this study was part of the district cohort of 14 TIF grant campuses, which were all chosen for participation in the grant because they had been historically under-performing and hard to staff. While leading those schools has its daily challenges, oftentimes administrators are inundated with well-meaning support that places huge demands on their time. With the introduction of a new observation rubric (FFT), more frequent teacher observations, and a new observation process (ID&E) that required pre- and post-conferences, all TIF administrators felt the strain of increased demands. The wide range of evidence statements reveals the lack of targeted training on how to gather, record, and report evidence. The state PDAS evaluation process, which is mandated at every Texas school, continued at the TIF campus alongside the ID&E observations. PDAS generally requires administrators to complete checklists more than it calls for them to record observation evidence to the extent required for ID&E observations. Additionally, conferences following a PDAS evaluation are optional and typically skipped. It is feasible to consider that some of the evidence recorded on the ID&E scorecards was limited and summary in nature, since this is outside the general procedure of observation, and conducted with an instrument that was unfamiliar to most prior to involvement in the TIF grant.

Limitations of the Study

Internal Validity

Causation conditions. This study was not designed to examine causation. However, it may be seen as a partial investigation on the conditions of causation. Since state standardized tests were given at the end of the school year, VAM data were generated after observation rubric data were collected. If a strong relationship had been found between the two measures in this study, a starting point might have been established to investigate in subsequent replications whether the level of teaching performances, as measured by the FFT are possible determinants of the level of student performances, as measured by VAM. This relationship, however, was not found in this study.

Instrumentation. VAM data were reported at the grade level/content area cohort level only. The district set the report policy. All mathematics teachers at a given grade level were given the same VAM score based on the progress of all tested students in the cohort. In some cases, there was a single teacher who taught all mathematics classes at a grade level, rendering the VAM cohort score a teacher-level score. In most cases, however, more than one teacher shared the same VAM score for the entire grade level cohort. VAM score attribution was not clear when two or more teachers teamed on a grade level.

This is problematic due to the potential inaccuracies incorporated into assigning V levels to teachers and then coupling that rating with T levels to then categorize teachers overall as high or low on both measures. While teacher level VAM data are susceptible to imprecisions of their own, based primarily on the fewer numbers of student scores factored into the calculation, the more direct attribution to individual teachers enhances studies such as this one. Beyond research uses, teacher level VAM data can be used as a powerful source of information for teachers to reflect on their practice and make adjustments.

External Validity

Population. Generalization of the study's results to wider populations is not reasonable due to the narrowness of its focus on only five elementary schools and one content area at three grade levels. The conditions that led to these schools being included in the district's TIF grant cohort, namely that they are historically hard to staff and under-performing, might also affect the ability to generalize from this study. However, due to the abundance of TIF grant projects using both FFT and VAM, it would seem logical that replication would be possible and valuable.

Accessible population. As previously noted, the ten school administrators who conducted the participated in this study had a variety of experience with mathematics instruction in Grades 3-5 and varying levels of experience using the FFT. Had there been more elementary schools using the FFT, the effect of the levels of experience could have been reduced on the findings, if, for example ten administrators with three years experience each using the rubric participated. Further, if all ten participants either had or had not served as elementary mathematics coaches, an analysis of observation data rule out or at least reduce observer experience as a contributing influence on the variation.

Reactivity. This was a twofold concern. First, according to local grant implementation policy, the two main observations that produced the ID&E scores were scheduled in advance. While observers had the leeway to include evidence from previous unscheduled observations, the scheduled observations formed the core of the evidence source. Teachers might have been inclined to perform and act in ways that were not representative of their typical daily instruction. This was exemplified in an evidence statement from an ID&E scorecard, in which the observing administrator noted, “Teacher attempts to maintain order in the classroom but has uneven success. Today's observation the students were on task.” Secondly, the ten participating administrators who took part in interviews about the ID&E observations, the evidence they noted, and the scores they gave, might also have been influenced by the awareness that they were participating in a study. One administrator announced before the first question was asked that she was not a ‘math person’ and did not feel as comfortable discussing mathematics as she might discussing literacy instruction. Several asked, after completing answers to questions, whether their responses were sufficient for the purposes of the interview. Both of these types of responses pointed to the possible effect I, as the researcher, might have had on the study.

The administrators’ awareness of my role as the TIF grant project manager and previous experience as a mathematics coach and content specialist not only might have influenced interview responses, but definitely impacted my interpretation of the data. I found myself in early drafts of this dissertation focusing on the findings surrounding the high T/high V level teachers, perhaps because that would have been an ideal outcome from the grant implementation. When the correlation turned out to be modest and not statistically significant, I understood that one of the underlying assumptions of the grant was not supported. With the help of my dissertation chair, I refocused on the story, in fact, that the data were telling, not on what I thought they would, or even should, tell. This helped me to represent all of the findings more completely and equitably.

Additionally, my background and experience as both a former teacher of elementary mathematics and as a content specialist and coach of elementary mathematics affected my role as researcher. Particularly with the interview data, I found myself unintentionally screening responses for terms and concepts that were content-specific. I initially I focused more on responses that incorporated mathematics-specific explanations. Later, however, I was able to apply my expertise by assigning mathematics concepts and terminology to those responses expressed without content-specific references.

Implications of the Study

This study can contribute to the conversation on measuring teacher effectiveness, particularly within the ranks of upper elementary mathematics instruction. If classroom observation rubric scores of teacher performance correlate to VAM scores of student progress, and if a set of best practices that are characteristic of teachers rated high in both measures can be determined, a plausible coaching plan might be designed and developed based on those practices.

According to ID&E rubric evidence and administrator interviews the high T/high V teachers excelled instructionally in developing conceptual understanding with their students by continually moving from concrete models to visual representations to the abstract level with explicit support, and insisting that students talk in detail about what they understand. For example, one administrator explained in an interview, “Manipulatives were used during small group to teach the students division conceptually, and then graph paper was used to assist students when setting up their division problems because some students have problems with that.” Scorecard evidence noted, “Students were able to explain relationship between multiplication and division; students feel safe and take risks engaging in mathematical discourse, even at basic levels.” These are examples from this relatively small study as starting points for an inventory of effective practices.

Theoretical Implications

This study supports the position that teaching effectiveness is not easily quantifiable and is very much open to interpretation. While quantitative measures such as VAM hold appeal for what they purport to capture, namely dispassionate evidence of instructional effectiveness, teaching and learning are processes that result from complex human interactions not easily or fully characterized by numbers. Qualitative investigations are vital counterpoints that can potentially explain the processes and interactions that might have impacted those numbers. As noted previously in chapter 1, Erickson (1986) asserted that the vast and subtle interconnectedness of classroom life is chiefly, but not entirely, constructed by the teacher. Observing teaching practice, making sense of it, and ultimately judging it is a difficult process, highly dependent on the skill, experience, and training of observers. Researching the complexities of this process can benefit from a post-positivist theoretical approach, as was applied to this study, whose fundamental epistemological assumption is that knowledge is constructed through the interpretation of experiences. Because of that, this study has shown that observation evidence reveals as much about the observers as it does about those who were observed.

Implications for Practice

The goal of the ID&E process was to hold meaningful discussions on instructional practice, based on classroom observations and expressed in the language of the FFT, so that actionable next steps can be taken to continuously support and develop teaching. The process was conceived as formative and ongoing. With fairly consistent results across the five campuses at every T level (Table 7 and Figure 2, chapter 4) regarding the strongest and weakest FFT rubric components, next steps are clearly presented regarding which specific components can be improved through targeted professional development and instructional coaching. High-yield classroom practices were confirmed in administrator interviews as well. An effective observation process that intentionally closes the feedback loop by defining the content, frequency, and manner of coaching opportunities to improve practice can have a positive impact on student progress and achievement.

Implications for Policy

VAM data have been integrated into the decision making process with high-stakes implications regarding employment, placement, and compensation. The results of this study can contribute to the debate about the appropriateness of using VAM data for such purposes. While this study used cohort-level VAM data, which has its own drawbacks when ascribed to individual teachers, even the use of teacher-level VAM data is problematic. Goldhaber and Hansen (2012), Sass (2008), and Amrein-Beardsley (2009), have shown that teacher-level VAM data can vary widely in consecutive years and not become reliably steady until multiple years have been collected. Basing high stakes decisions on teachers' VAM data from just one or two years should be approached with some caution. While multiple years' data are preferred, questions remain about how much weight VAM data should carry in influencing those decisions.

Recommendations for Future Researchers

Future research on the relationship between observation rubric scores and student progress measures would best be served by accessing and using teacher-level VAM data. Even though there are concerns about VAM's volatility, and the non-specificity of the FFT, the attribution is stronger than when using cohort-level data. Had those data been accessible in this study, a stronger relationship between the two measures might have been established.

A second recommendation is to observe the pre and post conferences that framed the ID&E observation. While it was noted that some ID&E scorecards had very general evidence statements recorded by administrators, it might be that those were merely notes that would be elaborated upon during a richer, more thorough discussion in the post conference about the observed lesson.

Third, accompanying administrators on classroom observations might provide a context to understand what they saw and heard, what they noted as evidence, and how they coded that evidence. The assumption that administrators accurately capture all of the most important facets that contribute to teaching and learning in an observed lesson may not be entirely true.

Fourth, additional content-focused training might benefit the entire observational process by cultivating all administrators' abilities to recognize the details of effective content specific instruction and provide teachers with more targeted feedback. As Nelson and Sassi (2007) noted, "Focusing on students' mathematical understanding and on teachers' ability to understand that thinking and interact with it, is a different classroom observation focus for many principals" (p. 56). Just as student compliance is often mistaken for engagement or the presence of math manipulatives is mistaken for the purposeful use of them to help students solve problems, the asking of questions and the occurrence of classroom discussions might also be mistaken for effective questioning and enriching discussions that benefit student understanding. Park, Nava, and Applegate (2011) formulated a mathematics observation rubric that makes mathematics discourse one of its four domains, emphasizing just how crucial questioning and discussions are in mathematics. The domain is comprised of two teacher discourse components – questioning and linking ideas, two student discourse components – linking ideas and mathematics rigor, student participation, and participation structures. Considering the depth and subtlety devoted to describing mathematical discourse in these terms might positively supplement the implementation of the FFT, where discourse is considered, generally, as one of the five instructional domain components.

Finally, it may be useful to expand the scope of a future study to include several years of VAM data and observation rubric data for teachers, along with multiple interviews with administrators. This could solidify findings and strengthen emergent patterns and themes in the data.

Conclusions

Measuring teaching effectiveness is a complex process. While one metric (quantitative or qualitative) cannot fully capture and characterize effective teaching, balancing multiple metrics is also a difficult, yet necessary, undertaking. The two measures of effective teaching examined in this study – teacher instructional performance as measured by observation rubric scores using the FFT and student growth as measured by VAM mean NCE gains using state standardized tests – were assumed to have a strong relationship. The findings in this study showed only a moderate relationship between the two measures that was not statistically significant. This is not to suggest, however, that they are unrelated: more than half of the 28 teachers studied earned consistent rankings on both measures – low/low or high/high.

The ten campus administrators who conducted the ID&E observations at the five elementary schools studied varied in the ways they captured, cited, and explained observation evidence. The depth, detail, and focus of evidence were seen when examining the ID&E scorecards. Whether evidence was scant or prolific, general or specific, patterns emerged suggesting agreement on what counted as evidence of effective practice for each component. Further, during interviews, trends and commonalities emerged in the explanations administrators provided on what comprises both effective and ineffective mathematics instruction. While experience levels with teaching and coaching mathematics differed among the administrators, responses were not as disparate as those differences in experience might initially suggest. Those with more direct experience with mathematics instruction in Grades 3-5 might have more readily recognized effective mathematics instruction and articulated their observations with greater specificity, but all administrators were able to use the language of the FFT to describe teachers' practice. As discussed above, perhaps integrating content-specific observation training might benefit the entire observation process when a rubric such as the FFT is used.

Results of this study support the need for replication in other TIF settings or similar settings where standardized testing data are used to determine teaching effectiveness. VAM, in particular, is a highly topical data source due to its recent proliferation and use in school districts across the country, and its inclusion as a required measure in federal grant programs. Before VAM data are used in high stakes educational decision-making, more ongoing research is necessary to determine the extent to which VAM relate to, not only observation rubric scores of teachers, but other measures of teaching effectiveness as well. Ultimately, further research might thoughtfully inform how multiple measures of teaching should be incorporated and weighted into a multi-faceted teacher evaluation system.

APPENDIX A

ID&E PRE-CONFERENCE FORM

Teacher ID&E Pre-Conference Form

Teacher ________________________ School __________________

Grade Level(s) ___________________ Subject(s) _______________

Observer ________________________ Date ____________________

Questions for discussion:

1. What are your learning outcomes for this lesson?

2. How have your instructional choices been influenced by your understanding of your students' academic strengths and weaknesses?

3. How will you engage all of your students in the learning, including those with special needs?

4. What type of grouping arrangements will you use during the lesson?

5. How and when will you know whether the students have learned what you intend?

APPENDIX B

ID&E POST-CONFERENCE FORM

Teacher ID&E Post-Conference Form

Teacher ______________________________ School ____________________

Grade Level(s) _______________________ Subject(s) __________________

Observer _____________________________ Date ______________________

Questions for Discussion:

1. In general, how successful was the lesson? To what extent did the students learn what you intended for them to learn? How do you know?

2. Comment on your classroom procedures, student conduct, and your use of physical space. To what extent did these contribute to student learning?

3. Did you depart from your plan? If so, how and why?

4. If you had an opportunity to teach this lesson again to the same group of students, what would you do differently?

APPENDIX C

ID&E SCORECARD TEMPLATE

TIF TEACHER INDIVIDUAL DEVELOPMENT & EVALUATION SCORECARD

Teacher: Observed By:

Class Information: Observation Date:

Start Time: End Time:

This form is intended for schools using the Teacher ID&E Scorecard adapted from the Danielson Framework for teaching. It is to be used only by campus administrators and teachers that have been trained in its use.

1e Designing Coherent Instruction

Component 1e is comprised of four elements.

1. Learning activities.
2. Instructional material and resources.
3. Instructional groups.
4. Lesson and unit structure.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 1e

2a Creating an Environment of Respect and Rapport

Component 2a is comprised of two elements.

1. Teacher interactions with students.

2. Student interactions with each other.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 2a

2b Establishing a Culture for Learning

Component 2b is comprised of three elements.

1. Importance of the content and learning.

2. Expectations for learning and achievement.

3. Student pride in work.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 2b

2c Managing Classroom Procedures

Component 2c is comprised of four elements.

1. Management of instructional groups.

2. Management of transitions.

3. Management of materials and supplies.

4. Performance of non-instructional duties.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 2c

2d Managing Student Behavior

Component 2d is comprised of three elements.

1. Expectations.

2. Monitoring student behavior.

3. Response to student misbehavior.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 2d

Component 2e Organizing Physical Space

Component 2e is comprised of two elements.

1. Safety and accessibility.

2. Arrangement of resources and use of physical resources, including technology when appropriate.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 2e

3a Communicating with Students

Component 3a is comprised of four elements.

1. Expectations for learning.

2. Directions and procedures.

3. Explanation of content.

4. Use of oral and written language.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 3a

3b Using Questioning and Discussion Techniques

Component 3b is comprised of three elements.

1. Quality of questions/prompts.
2. Discussion techniques.
3. Student participation.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 3b

3c Engaging Students in Learning

Component 3c is comprised of four elements.

1. Activities and assignments.

2. Grouping of students.

3. Instructional materials and resources.

4. Structure and pacing.

Unsatisfactory	Basic	Proficient	Distinguished

Evidence for Component 3c

3d Using Assessment in Instruction

Component 3d is comprised of four elements.

1. Assessment criteria.

2. Monitoring of student learning.

3. Feedback to students.

4. Student self-assessment and monitoring of progress.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 3d

3e Demonstrating Flexibility and Responsiveness

Component 3e is comprised of three elements.

1. Lesson adjustment.

2. Response to students.

3. Persistence.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 3e

4a Reflecting on Teaching

Component 4a is comprised of two elements.

1. Accuracy.

2. Use in future teaching.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 4a

4b Maintaining Accurate Records

Component 4b is comprised of three elements.

1. Student completion of assignments.

2. Student progress in learning.

3. Non-instructional records.

Unsatisfactory Basic Proficient Distinguished

Evidence for Component 4b

General Comments/Suggestions

APPENDIX D

ID&E SCORECARD RUBRIC

From the Framework for Teaching Evaluation Instrument (Danielson, 2011), accessed through the district licensed account for the Teachscape website.

Teacher Individual Development and Evaluation (ID&E) Scorecard Rubric

Levels of Performance and Rubric Scores

Level 4: Distinguished	4 points
Level 3: Proficient	3 points

Level 2: Basic	2 points
Level 1: Unsatisfactory	0 points

Component 1e:

Designing Coherent Instruction

Distinguished

The sequence of learning activities follows a coherent sequence, is aligned to instructional goals, and is designed to engage students in high-level cognitive activity. These are appropriately differentiated for individual learners. Instructional groups are varied appropriately, with some opportunity for student choice.

Proficient

Most of the learning activities are aligned with the instructional outcomes and follow an organized progression suitable to groups of students. The learning activities have reasonable time allocations; they represent significant cognitive challenge, with some differentiation for different groups of students and varied use of instructional groups.

Basic

Some of the learning activities and materials are aligned with the instructional outcomes and represent moderate cognitive challenge, but with no differentiation for different students. Instructional groups partially support the activities, with some variety.

The lesson or unit has a recognizable structure; but the progression of activities is uneven, with only some reasonable time allocations.

Unsatisfactory

Learning activities are poorly aligned with the instructional outcomes, do not follow an organized progression, are not designed to engage students in active intellectual activity, and have unrealistic time allocations. Instructional groups are not suitable to the activities and offer no variety.

Component 2a:

Creating an Environment of Respect and Rapport

Distinguished

Classroom interactions among the teacher and individual students are highly respectful, reflecting genuine warmth and caring and sensitivity to students as individuals. Students exhibit respect for the teacher and contribute to high levels of civility among all members of the class. The net result of interactions is that of connections with students as individuals.

Proficient

Teacher-student interactions are friendly and demonstrate general caring and respect. Such interactions are appropriate to the ages of the students. Students exhibit respect for the teacher. Interactions among students are generally polite and respectful. The teacher responds successfully to disrespectful behavior among students. The net result of the interactions is polite and respectful, but business-like.

Basic

Patterns of classroom interactions, both between the teacher and students and among students, are generally appropriate but may reflect occasional inconsistencies, favoritism, and disregard for students' ages, cultures, and developmental levels. Students rarely demonstrate disrespect for one another. The teacher attempts to respond to disrespectful behavior, with uneven results. The net result of the interactions is neutral: conveying neither warmth nor conflict.

Unsatisfactory

Patterns of classroom interactions, both between the teacher and students and among students, are mostly negative, inappropriate, or insensitive to students' ages, cultural backgrounds, and developmental levels. Interactions are characterized by sarcasm, put-downs, or conflict. The teacher does not deal with disrespectful behavior.

Component 2b:

Establishing a Culture for Learning

Distinguished

The classroom culture is a cognitively vibrant place, characterized by a shared belief in the importance of learning. The teacher conveys high expectations for learning by all students and insists on hard work; students assume responsibility for high quality by initiating improvements, making revisions, adding detail, and/or helping peers.

Proficient

The classroom culture is a cognitively busy place where learning is valued by all, with high expectations for learning the norm for most students. The teacher conveys that with hard work students can be successful; students understand their role as learners and consistently expend effort to learn. Classroom interactions support learning and hard work.

Basic

The classroom culture is characterized by little commitment to learning by the teacher or students. The teacher appears to be only "going through the motions," and students indicate that they are interested in completion of a task rather than quality. The teacher conveys that student success is the result of natural ability rather than hard work; high expectations for learning are reserved for those students thought to have a natural aptitude for the subject.

Unsatisfactory

The classroom culture is characterized by a lack of teacher or student commitment to learning and/or little or no investment of student energy in the task at hand. Hard work is not expected or valued. Medium to low expectations for student achievement are the norm, with high expectations for learning reserved for only one or two students.

Component 2c:

Managing Classroom Procedures

<u>Distinguished</u>

Instructional time is maximized due to efficient classroom routines and procedures. Students con- tribute to the management of instructional groups, transitions, and/or the handling of materials and supplies. Routines are well understood and may be initiated by students.

<u>Proficient</u>

There is little loss of instructional time due to effective classroom routines and procedures. The teacher's management of instructional groups and/or the handling of materials and supplies is consistently successful. With minimal guidance and prompting, students follow established classroom routines.

<u>Basic</u>

Some instructional time is lost due to only partially effective classroom routines and procedures. The teacher's management of instructional groups, transitions, and/or the handling of materials and supplies is inconsistent, leading to some disruption of learning. With regular guidance and prompting, students follow established routines.

Unsatisfactory

Much instructional time is lost due to inefficient classroom routines and procedures. There is little or no evidence of the teacher managing instructional groups, transitions, and/or the handling of materials and supplies effectively. There is little evidence that students know or follow established routines.

Component 2d:

Managing Student Behavior

Distinguished

Student behavior is entirely appropriate. Students take an active role in monitoring their own behavior and that of other students against standards of conduct. The teacher's monitoring of student behavior is subtle and preventive. The teacher's response to student misbehavior is sensitive to individual student needs and respects student dignity.

Proficient

Student behavior is generally appropriate. The teacher monitors student behavior against established standards of conduct. The teacher 's response to student misbehavior is consistent, appropriate and respectful to students, and effective.

Basic

Standards of conduct appear to have been established, but their implementation is inconsistent. The teacher tries, with uneven results, to monitor student behavior and respond to student misbehavior. There is inconsistent implementation of the standards of conduct.

Unsatisfactory

There appear to be no established standards of conduct and little or no teacher monitoring of student behavior. Students challenge the standards of conduct. Response to student misbehavior is repressive, or disrespectful of student dignity.

Component 2e:

Organizing Physical Space

Distinguished

The classroom is safe, and the physical environment ensures the learning of all students, including those with special needs. Students contribute to the use or adaptation of the physical environment to advance learning. Technology is used skillfully, as appropriate to the lesson.

Proficient

The classroom is safe, and learning is accessible to all students; teacher ensures that the physical arrangement is appropriate to the learning activities. Teacher makes effective use of physical resources, including computer technology.

Basic

The classroom is safe, and essential learning is accessible to most students, and the teacher's use of physical resources, including computer technology, is moderately effective. Teacher may attempt to modify the physical arrangement to suit learning activities, with partial success.

Unsatisfactory

The physical environment is unsafe, or some students don't have access to learning.

There is poor alignment between the physical arrangement and the lesson activities.

Component 3a:

Communicating with Students

Distinguished

The teacher links the instructional purpose of the lesson to student interests; the directions and procedures are clear and anticipate possible student misunderstanding. The teacher's explanation of content is thorough and clear, developing conceptual understanding through artful scaffolding and connecting with student interests. Students contribute to extending the content and explaining concepts to their classmates. The teacher's spoken and written language is expressive, and the teacher finds opportunities to extend students' vocabularies.

Proficient

The instructional purpose of the lesson is clearly communicated to students, including where it is situated within broader learning; directions and procedures are explained clearly. The teacher's ex- planation of content is well scaffolded, clear, and accurate, and connects with student knowledge and experience. During the explanation of content, the teacher invites student intellectual engagement. The teacher's spoken and written language is clear and correct. Vocabulary is appropriate to students' ages and interests.

Basic

The teacher's attempt to explain the instructional purpose has only limited success, and/or directions and procedures must be clarified after initial student confusion. The teacher's explanation of the content may contain minor errors; some portions are clear while other portions are difficult to follow. The teacher's explanation consists of a monologue, with no invitation to the students for intellectual engagement. The teacher 's spoken language is correct; however, vocabulary is limited or not fully appropriate to students' ages or backgrounds.

Unsatisfactory

The instructional purpose of the lesson is unclear to students and the directions and procedures are confusing. The teacher's explanation of the content contains major errors. The teacher's spoken or written language contains errors of grammar or syntax. Vocabulary is inappropriate, vague, or used incorrectly, leaving students confused.

Component 3b:

Using Questioning and Discussion Techniques

Distinguished

The teacher uses a variety or series of questions or prompts to challenge students cognitively, advance high-level thinking and discourse, and promote meta-cognition. Students formulate many questions, initiate topics and make unsolicited contributions. Students themselves ensure that all voices are heard in the discussion.

Proficient

While the teacher may use some low-level questions, he or she poses questions to students designed to promote student thinking and understanding. The teacher creates a genuine discussion among students, providing adequate time for students to respond and stepping aside when appropriate. The teacher successfully engages most students in the discussion, employing a range of strategies to ensure that most students are heard.

Basic

The teacher's questions lead students along a single path of inquiry, with answers seemingly deter- mined in advance. Or, the teacher attempts to frame some questions designed to promote student thinking and understanding, but only a few students are involved. The teacher attempts to engage all students in the discussion and to encourage them to respond to one another, with uneven results.

Unsatisfactory

The teacher's questions are of low cognitive challenge, with single correct responses, and asked in rapid succession. Interaction between teacher and students is predominantly recitation style, with the teacher mediating all questions and answers. A few students dominate the discussion.

Component 3c:

Engaging Students in Learning

Distinguished

Virtually all students are intellectually engaged in challenging content through well designed learning tasks and suitable scaffolding by the teacher. Learning tasks and activities are fully aligned with the instructional outcomes. In addition, there is evidence of some student initiation of inquiry and student contributions to the exploration of important content. The lesson has a clearly defined structure, and the pacing of the lesson provides students the time needed to intellectually engage with and reflect upon their learning, and to consolidate their understanding. Students may have some choice in how they complete tasks and may serve as resources for one another.

Proficient

The learning tasks and activities are aligned with the instructional outcomes and are designed to challenge student thinking, resulting in active intellectual engagement by most students with important and challenging content, and with teacher scaffolding to support that engagement. The lesson has a clearly defined structure and the pacing of the lesson is appropriate, providing most students the time needed to be intellectually engaged.

Basic

The learning tasks and activities are partially aligned with the instructional outcomes but require only minimal thinking by students, allowing most students to be passive or merely compliant. The lesson has a recognizable structure; however the pacing of the lesson may not provide students the time needed to be intellectually engaged.

Unsatisfactory

The learning tasks and activities, materials, resources, instructional groups, and technology are poorly aligned with the instructional outcomes, or require only rote responses. The lesson has no clearly defined structure, or the pace of the lesson is too slow or rushed. Few students are intellectually engaged or interested.

Component 3d:

Using Assessment in Instruction

Distinguished

Assessment is fully integrated into instruction through extensive use of formative assessment. Students appear to be aware of, and there is some evidence that they have contributed to, the assessment criteria. Students self-assess and monitor their progress. A variety of feedback, from both the teacher and peers, is accurate and specific and advances learning. Questions/prompts/ assessments are used regularly to diagnose evidence of learning, and instruction is adjusted and differentiated to address individual student misunderstandings.

Proficient

Assessment is regularly used during instruction through teacher and/or student monitoring of progress of learning, resulting in accurate, specific feedback that advances learning. Students appear to be aware of the assessment criteria; some of them engage in self-assessment. Questions/prompts/assessments are used to diagnose learning, and adjustment to instruction is made to ad- dress student misunderstandings.

Basic

Assessment is sporadically used to support instruction through some teacher and/or student monitoring of progress of learning. Feedback to students is general, and students are only partially aware of the assessment criteria; few assess their own work. Questions/prompts/assessments are rarely used to diagnose evidence of learning. Adjustment of the lesson in response to the assessment is minimal or ineffective.

Unsatisfactory

There is little or no assessment or monitoring of student learning; feedback is absent or of poor quality. Students do not appear to be aware of the assessment criteria and do not engage in self-assessment. There is no attempt to adjust the lesson as a result of assessment.

Component 3e:

Demonstrating Flexibility and Responsiveness

Distinguished

Teacher seizes opportunities to enhance learning, building on a spontaneous event or student interests. Teacher ensures the success of all students, using an extensive repertoire of instructional strategies.

Proficient

Teacher promotes the successful learning of all students, making adjustments as needed to instruction plans and accommodating student questions, needs and interests.

Basic

Teacher attempts to modify the lesson when needed and to respond to student questions, with moderate success. Teacher accepts responsibility for student success, but has only a limited repertoire of strategies to draw upon.

Unsatisfactory

Teacher adheres to the instruction plan, even when a change would improve the lesson or of students' lack of interest. Teacher brushes aside student questions; when students experience difficulty, the teacher blames the students or their home environment.

Component 4a:

Reflecting on Teaching

Distinguished

The teacher makes a thoughtful and accurate assessment of a lesson's effectiveness and the extent to which it achieved its instructional outcomes, citing many specific examples from the lesson and weighing the relative strengths of each. Drawing on an extensive repertoire of skills, the teacher offers specific alternative actions, complete with the probable success of different courses of action.

Proficient

The teacher makes an accurate assessment of a lesson's effectiveness and the extent to which it achieved its instructional outcomes and can cite general references to support the judgment. The teacher makes a few specific suggestions of what could be tried another time the lesson is taught.

Basic

The teacher has a generally accurate impression of a lesson's effectiveness and the extent to which instructional outcomes were met. The teacher makes general suggestions about how a lesson could be improved.

Unsatisfactory

The teacher does not know whether a lesson was effective or achieved its instructional outcomes, or the teacher profoundly misjudges the success of a lesson. The teacher has no suggestions for how a lesson could be improved.

Component 4b:

Maintaining Accurate Records

Distinguished

The teacher's system for maintaining information on student completion of assignments, student progress in learning, and non-instructional records is fully effective. Students contribute information and participate in maintaining the records.

Proficient

The teacher's system for maintaining information on student completion of assignments, student progress in learning, and non-instructional records is fully effective.

Basic

The teacher's system for maintaining information on student completion of assignments and student progress in learning is rudimentary and only partially effective. The teacher's records for non-instructional activities are adequate, but inefficient, and, unless given frequent oversight, are prone to errors.

Unsatisfactory

The teacher's system for maintaining information on student completion of assignments and student progress in learning is nonexistent or in disarray. The teacher's records for non-instructional activities are in disarray, the result being errors and confusion.

APPENDIX E

ADMINISTRATOR INTERVIEW QUESTIONS

Administrator Interview Questions

What do you typically find in a mathematics classroom where the lesson seems to be working well?

In the most effective mathematics lessons you observed, which ID&E rubric components did you rate the highest and why?

In the least effective mathematics lessons you observed, which ID&E rubric components did you rate the lowest and why?

Which rubric components do you think best capture effective mathematics instruction?

Is there a particular FFT (rubric) component that you see as an area of need to improve mathematics instruction at your campus?

REFERENCES

Akiba, M., & Letendre, G. (2009). *Improving teacher quality: The U.S. teaching force in global context*. New York, NY: Teachers College Press.

Alvarez, M. E., & Anderson-Ketchmark, C. (2011). Danielson's framework for teaching. *Children & Schools, 33*(1), 61.

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65-75.

Amrein-Beardsley, A. (2009). Value-added tests: Buyer, beware. *Educational Leadership*, 38-42.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS educational value-added assessment system in the Houston independent school district: Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12), 1-31.

Ball, D. L., & Forzani, F. M. (2010). What does it take to make a good teacher? *Kappan, 92*(2), 8-12.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Banchero, S., & Kesmodel, D. (2011, Sept. 13). Teachers are put to the test: More states tie tenure, bonuses to new formulas for measuing student scores, *Wall Street Journal*.

Battelle for Kids Roster Verification. (2014). Retrieved from http://www.battelleforkids.org/how-we-help/strategic-measures/data-services-roster-verification.

Boston, M. (2012). Assessing instructional quality in mathematics. *The Elementary School Journal, 113*(1), 76-104. doi: 10.1086/666387

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. *ETS Policy Information Perspective*, 1-16.

Briggs, D. C., & Weeks, J. P. (2011). The persistence of school-level value-added. *Journal of Educational and Behavioral Statistics, 36*, 616-637.

Burns, M. (2000). *About teaching mathematics* (2nd ed.). Sausalito, CA: Math Solutions.

Cantrell, S., & Scantlebury, J. (2011). Effective teaching: What is it and how is it measured? *Voices in Urban Education*, *31*, 28-35.

Capraro, M. M. , Capraro, R. M., Carter, T., & Harbaugh, A. (2010). Understanding, questioning, and representing mathematics: What makes a difference in middle school classrooms? *Research in Middle Level Education, 34*(4), 1-19.

Chapin, S. H , & Johnson, A. (2006). *Math matters: Understanding the math you teach*. Sausalito, CA: Math Solutions.

Charmaz, K. (2010). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: ASCD.

Danielson, C. (2011). *The framework for teaching evaluation instrument*. Alexandria, VA: ASCD.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: ASCD.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15.

Di Carlo, M. (2012). How to use value-added measures right. *Educational Leadership*, 38-42.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119-161). New York, NY: MacMillan Press.

Ewing, J. (2011). Mathematical intimidation: Driven by the data. *Notices of the American Mathematical Society, 58*(5), 667-673.

Ferrao, E. (2012). On the stability of value added indicators. *Qual Quant, 46*, 627-637. doi: 10.1007/s11135-010-9417-6

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher, 43*, 100-107.

Floden, R. E. (2012). Teacher value added as a measure of program quality: Interpret with caution. *Journal of Teacher Education, 63*(5), 356-360.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Retrieved from http://www.brookings.edu/research/reports/2010/11/17-evaluating-teachers.

Glesne, C. (2011). *Becoming qualitative researchers: An introduction* (4th ed.). Boston, MA: Pearson Education, Inc.

Goldhaber, D., & Hansen, M. (2012). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *CALDER Working Paper No. 73*, 1-42.

Grbich, C. (2010). *Qualitative data analysis: An introduction*. Thousand Oaks, CA: SAGE Publications, Inc.

Hannaway, J., & Mittelman, J. (2011). Education politics and policy in the era of evidence. In D. E. Mitchell, R. L. Crowson, & D. Shipps (Eds.),*Shaping education policy: Power and process* (pp. 81-91). New York, NY: Routledge.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*, 466-479.

Hanushek, E. A., & Rivkin, S. G. (2010a). Generalizations about using value-added measures of teacher quality. *American Economic Review: Papers and Proceedings, 100*, 267-271.

Hanushek, E. A., & Rivkin, S. G. (2010b). The quality and distribution of teachers under the no child left behind act. *Journal of Economic Perspectives, 24*(3), 133-150.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 692-712. doi: 10.1002/pam.20462

Harris, D. N. (2010). Clear away the smoke and mirrors of value-added. *Phi Delta Kappan, 91*(8), 66-69.

Henderson, K.A. (2011). Post-positivism and the pragmatics of leisure research. *Leisure Sciences, 33*, 341-346. doi: 10.1080/01490400.2011.583166

Hill, H. C., Blunk, M. L., Charalombous, Y. C., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430-511.

Hill, H. C., Kapitula, L. R., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. doi: 10.3102/0002831210387916

Hill, H. C., Rowan, B, & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406. doi: 10.3102/00028312042002371

Innes, R. G. (2010). *KERA (1990-2010): What have we learned?* Bluegrass Institute for Public Policy Solutions.

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. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kennedy, M. M. (2006). From teacher quality to quailty teaching. *Educational Leadership, 63*(6), 14-19.

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34-70.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in washoe county. *Peabody Journal of Education, 79*(4), 54-78.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review, 28*, 682-692.

Koretz, D., & Barron, S. (1998). The validity of gains in scores on the kentucky instructional results information system (KIRIS). RAND Education: RAND.

Kuhn, T. (1970). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the tennessee value added assessment system. *Educational Evaluation and Policy Analysis, 25*(3), 287-298.

Lasley II, T. J., Siedentop, D., & Yinger, R. (2006). A systemic approach to enhancing teacher quality: The Ohio model. *Journal of Teacher Education, 57*(1), 13-21. doi: 10.1177/0022487105284455

Lincove, J. A., Osborne, C., Dillon, A., & Mills, N. (2014). The politics and statistics of value-added modeling for accountability of teacher preparation programs. *Journal of Teacher Education, 65*(1), 24-38.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability. *Journal of Educational and Behavioral Statistics, 27*(3), 255-270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67. doi: 10.1111/j.1745-3984.2007.00026.x

Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement. *European Journal of Education Research, Development and Policy, 46*(4), 440-455.

Malen, B. (2011). An enduring issue: The relationship between political democracy and educational effectiveness. In D. E. Mitchell, R. L. Crowson, & D. Shipps (Eds.), *Shaping education policy: Power and process* (pp. 23-60). New York, NY: Routledge.

Marshall, K. (2009). *Rethinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap*. San Francisco, CA: Jossey-Bass.

Matula, J. J. (2011). Embedding due process measures throughout the evaluation of teachers. *NASSP Bulletin, 95*(2), 99-121.

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.

Meyer, R. H., & Dokumaci, E. (2009). *Value-added models and the next generation of assessments*. Paper presented at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, Princeton, NJ.

Milanowski, A. T. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*, 33-53. doi: 10.1207/s15327930pje7904_3

Milanowski, A. T. (2005). Split roles in performance evaluation - A field study involving new teachers. *Journal of Personnel Evaluation in Education, 18*, 153-169.

Milanowski, A. T. (2011). Strategic measures of teacher performance. *Kappan, 92*(7), 19-25.

Nelson, B. S., & Sassi, A. (2007). What math teachers need most. *The Education Digest, 72*(6), 54-56.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis, 18*(23), 1-23.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi: 10.3102/0002831210362589

Park, J., Nava, I., & Appelgate, M. (2011). Observation rubric for secondary mathematics *Center Xchange*. Los Angeles, CA: UC Regents.

Paufler, N., & Amrein-Beardsley, A. (2013). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, 1-35. doi: 10.3102/0002831213508299

Phillips, K. J. R. (2010). What does "highly qualified" mean for student achievement? Evaluating the relationships between teacher quality indicators and at-risk students' mathematics and reading achievement gains in first grade. *The Elementary School Journal, 110*(4), 464-493.

Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education, 63*(5), 318-334.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Prince, C. D., Koppich, J., Azar, T. M., Bhatt, M., & Witham, P. J. (2010). Research synthesis: Measurement. *Center for Educator Compensation Reform*, 1-6.

Race to the Top. (2012). Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

Rivkin, S. G. (2007). Value-added analysis and education policy. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review, 100*(2), 261-266.

Sanders, W. L. (2003). *Beyond no child left behind*. Paper presented at the Annual Meeting American Educational Research Association, Chicago, IL.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247-256.

Sass, T. R. (2008). The stability of value-added measures of teacher quality and implications for teacher compensation policy. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research.

Scherrer, J. (2012). What's the value of VAM (value-added modeling)? *Phi Delta Kappan, 93*(8), 58-60.

Schlechty, P. C. (2011). *Engaging students: The next Level of working on the work*. San Francisco, CA: Jossey-Bass.

Schmoker, M. (2012). The madness of teacher evaluation frameworks. *Kappan, 93*(8), 70-71.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355. doi: 10.1177/0022487111404241=20

Teacher Incentive Fund. (2012). Retrieved from http://www2.ed.gov/programs/teacherincentive/index.html

Teddlie, C. & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13*(1), 12-28.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: SAGE Publications, Inc.

Towndrow, P. A., & Tan, K. (2009). Teacher self-evaluation and power. *Teacher Development, 13*(3), 285-292.

TVAAS: An introduction to value-added in Tennessee. (2012, June) *Taking Note*: State Collaborative on Reforming Education.

TVAAS Resources. (2012). Retrieved from http://www.tn.gov/education/data/TVAAS.shtml

U.S. Congress. (2001). No Child Left Behind Act of 2001.

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*.

Weston, S. P., & Sexton, R. F. (2009). *Substantial and yet not sufficient: Kentucky's effort to build proficiency for each and every child*. Paper presented at the Campaign for Educational Equity, Teachers College Columbia University.

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.

Yeh, S. S. (2012). The reliability, impact, and cost-effectiveness of value-added teacher assessment methods. *Journal of Education Finance, 37*(4), 374-399.

Young, D. C. (2009). Interpretivism and education law research: A natural fit. *Education Law Journal, 18*(3), 203-219.