JAEPR, 2(*2*), 36-55

ISSN: 1930-9325

RESEARCH ARTICLE

A Comparison of Classification Issues across Teacher Effectiveness Measures

Jon Brasfield

University of Findlay

In an educational landscape where teacher evaluation methods are increasingly discussed and scrutinized in research offices, legislatures, and school buildings, the differences in policy and instrumentation among states and school districts can paint a confusing picture of these varying methods and their impacts. To help assess the picture in North Carolina, this study examined teacher effectiveness data on 147 teachers from 16 schools in a large urban school district. Three measures of teacher effectiveness (a value-added measure of student growth, a third-party observation score, and state-mandated principal evaluations) were examined with a particular focus on how teachers were classified via the different methods. The research question examined the similarities and differences in classification across the measures. Correlational, cross-tabular, and agreement statistic results suggested that the value-added measure and the third-party observational measure were not independent. It was also found that principal ratings did little to differentiate between teachers.

Keywords: teacher evaluation, value-added modeling, teacher effectiveness, teacher observation

Today, more data on more aspects of education are available than ever before. The ability of states, districts, and educational research organizations to capture and link information on students and teachers in a longitudinal manner has far-reaching implications for the future of educational practice in the United States. One area in which this wealth of data has recently received attention both within education policy arenas and in mainstream society is the field of teacher effectiveness evaluation (e.g. Anderson, 2013; Banchero & Kesmodo, 2011; Ripley, 2010).

In the period from 2009-2012, 36 states and the District of Columbia made substantive changes to their policies regarding teacher evaluation, including guidelines on tenure, retention, dismissal, and in some cases, performance pay (National Council on Teacher Quality, 2013). Though many policy changes have been instituted, there remains much diversity in the policies themselves, as well as in the research regarding the best aspects of a quality teacher evaluation system.

In recent years, considerable research has been done on the efficacy of traditional methods of teacher evaluation, as well as on emerging methods of teacher evaluation such as value-added modeling (VAM), standards-based observation, and student surveys, among others. In 2011, the Bill and Melinda Gates Foundation completed a massive study on this topic in an attempt to determine the best methods of measuring teacher quality. Among their findings was that standards-based observations, student survey data, and value-added modeling of teacher effects on student learning could be valuable indicators of quality. Policy changes across states have been inspired by this research, with 39 states now requiring that classroom observations of teachers be incorporated into teachers' evaluations and 30 states requiring that some measure of student achievement be a significant or the most significant aspect of a teacher's evaluation (Collins & Amrein-Beardsley, 2013). According to a nationwide survey of educational effectiveness policy, 40 states and the District of Columbia currently are either using or developing some sort of growth measure or VAM (Collins & Amrein-Beardsley, 2013). In the past four years, the number of states requiring that student achievement play a role in teacher tenure decisions grew from zero to nine (NCTQ, 2013). In addition, where teachers previously were evaluated irregularly, 23 states now require that every teacher undergo annual evaluation.

With so much research being done on the effectiveness of various methods of measuring teacher effectiveness, opportunities arise for examination of the purpose of teacher evaluation. Papay (2012) proposed that evaluation systems serve two potential purposes: first, to identify high- and low-quality teachers for purposes of tenure, dismissal, or merit pay; and second, to serve as formative assessment for a teacher's practice, providing him or her with feedback crucial to the professional growth process. For a teacher evaluation system to serve either of these purposes, the conclusions reached by the use of these instruments must be meaningful and provide some diagnostic or prescriptive information. Different instruments used to sort teachers into performance-based groups would ideally sort teachers into the same groups, regardless of the influence of outside factors. In this study, three different teacher effectiveness measures from a set of elementary and middle schools in a large, urban school district in North Carolina were examined. The data were analyzed to assess how various teacher- and school-level variables affected measurements under various methods, as well as to determine if certain categories of teachers might be scored dissimilarly on various methods.

Teacher Evaluation in North Carolina

In North Carolina (at the time of data collection), teachers were subject to annual evaluation using the North Carolina Teacher Evaluation System (NC TES), a standards-based instrument designed to be completed by administrators. The system features five standards – one each regarding leadership, establishing a respectful environment for all students, content knowledge, facilitation of learning, and reflecting on their practice. Veteran teachers are evaluated annually on two standards of the NC TES instrument (leadership and facilitation of learning), and beginning teachers and teachers in review years are evaluated on all five standards In its recommendations, the NCTQ suggested that states use multiple measures of student learning to measure teacher effectiveness and require classroom observations that focus on and document the effectiveness of instruction, stating, "well-designed and executed observations provide the clearest opportunity to give teachers actionable feedback on the strengths and weaknesses of their instructional practice" (NCTQ, 2013, p.8).

Whereas the NC TES does require evidence of student learning and does require classroom observations, it does not track the fidelity of observations through records of inter-rater reliability. The data set used in this study involved a second set of standards-based observations conducted by full-time, highly trained observers who exhibited high inter-rater reliability – one trait not present in the literature on the NC TES. These observation scores are referred to as STAR, after the name of the grant program under which the observers were hired and trained. In addition to the NC TES, North Carolina teachers are also measured by a VAM, SAS Institute Inc.'s Education Value-Added Assessment System (EVAAS®).

The subset of schools used in this study is unique in that teachers are evaluated by three measures of teacher effectiveness instead of the typical two. North Carolina publishes aggregate data on NC TES outcomes (http://apps.schools.nc.gov/pls/apex/f?p=155:1), and statewide EVAAS® trends are available to educators in the state, but the addition of a highly-focused standards-based observation system to these schools allows an opportunity to observe the differences in levels of classification between the two existing methods of teacher measurement in the state side-by-side with another method that has shown promise in research.

To fully examine the relationships between the methods of teacher measurement used in this sample and the school- and teacher-level variables of interest, the following research question was proposed: Do the three methods of teacher effectiveness measurement (TEM) classify teachers in substantively different ways?

This question focused on the distributions created when teachers are classified solely on one of the three methods (NC TES, EVAAS, or STAR observation score). Analysis of these distributions allowed examination of the relationships between the methods. Statistical tests were conducted to determine if significant differences existed in the classification of individual teachers using different methods. Of particular interest were the relationships (or lack thereof) in teacher classification via the NC TES and classification via the separate standards-based observation rubric. Differences would suggest that two measures of classroom techniques are measuring different constructs in practice (or that one or both are unreliable or invalid), which ultimately may suggest that inclusion of a separate measure of classroom technique may be desirable.

REVIEW OF THE LITERATURE

Teacher Evaluation Systems

Evaluation by principals. Traditionally, teachers have been evaluated by a state- or district-mandated process that involves some degree of judgment based on observable behaviors, both in and outside the classroom. Teachers who score high on these evaluations, in general, tend to also score high on other measures of specific teacher effectiveness, such as classroom management and having better relationships with their students (Stronge, Ward, & Grant, 2011).

A significant amount of research has been conducted on these evaluative processes, and many of these studies have found that principals are the primarily responsible parties for conducting the evaluations (e.g., Zimmerman & Deckert-Pelton, 2003). Studies suggest further that principals are tasked with carrying out teacher evaluations without being well-trained in the process and with an inability to devote sufficient time to fidelity (Toch, 2008). Even when

fidelity is attained, the tools used for evaluation often don't focus directly on instructional quality (Toch, 2008).

Despite these challenges, "Principals' behaviors, expectations, and perceptions help build the climate of a school and these data suggest that teachers are looking to their principals to be leaders in this critical domain of assessment and evaluation" (Zimmerman & Deckert-Pelton, 2003, p. 35). The knowledge that the building administrators are the parties responsible for determining teacher effectiveness contributes to a sense of alertness and attention to quality by teachers, when the evaluations are perceived as useful (Noakes, 2008).

However, many teachers do not find these evaluations to be a positive experience. Multiple studies (Zimmerman & Deckert-Pelton, 2003; Arrington, 2010; McConney, Ayres, Hansen, & Cuthbertson, 2003; Van Tassel-Baska, Quek, & Feng, 2006) have found that teachers tend to respond poorly to the evaluation experiences carried out at the building level by their own administrators.

A constructive and collaborative relationship found to be lacking between principals and teachers can be established by administrators if they provide feedback which is perceived as honest and helpful (Marshall, 2005). To provide thoughtful feedback, principals must devote their full attention to a classroom for an uninterrupted block of time (Zatynski, 2012). If principals are to be the primary evaluators for teachers, then adequate preparation and ample time to establish a valid assessment of the teacher's performance in the classroom are crucial. Unfortunately, studies have found that this training and time allowance is not always adequate (Ingle, Rutledge, & Bishop, 2011; Johnson & Roellke, 1999; Marshall, 2005; Ramirez, Lamphere, Smith, Brown, & Pierceall-Herman, 2011).

Evaluation by third parties. If training and adequate observational time are obstacles to fair and accurate teacher evaluations, then a logical fix would be to assign the task of evaluations to outside observers who have both adequate training and time to devote to the fidelity of the process. The Measures of Effective teaching (MET) project was intended to build and test measures of effective teaching to find out how evaluation methods could best be used to help districts and teachers identify effective teaching and improve teacher quality. Over 3,000 teachers across six school districts volunteered to participate in the project, with the goal of determining how to "close the gap between expectations of effective teaching and what is actually happening in classrooms" (Measures of Effective Teaching Project, n.d.). The project examined student survey data, classroom observation data, student achievement data, teacher content knowledge, and teacher perceptions of working conditions.

As one aspect of their investigation, the MET project compared teacher evaluations conducted by principals to those conducted by "peers" – fellow teachers trained in and tasked with observing other teachers. In a sample of 129 administrators using a four-point observation scale, they found that principals rated their own teachers slightly higher (0.1 points) than principals from other schools and even slightly higher (0.2 points) than peers (Ho & Kane, 2013). These differences may seem quite small, but given the highly compressed data (scores were disproportionately clustered in the middle of the distribution, another common finding in principal evaluations), a 0.1 point difference in mean observation score was equivalent to ten percentile ranks.

If feedback from principal evaluations is to be useful, it would follow that teacher ratings should be differentiated among the rating categories. Studies have found that this is not always the case. One study addressing teacher evaluations of over 8,000 teachers across five large

school districts determined that almost 99 percent of teachers receive ratings of "satisfactory," based on their district's standards when a dichotomous rating system was used (Zatynski, 2012). A separate study showed that principals in 87 percent of schools in a single large urban district of over 600 schools did not issue a single "unsatisfactory" rating between 2003 and 2006 (Toch, 2008).

The MET project found a similar phenomenon in a different district:

> In a study where 127 school administrators and teachers observers each conducted 24 teacher observations, observers rarely used the top or bottom categories ("unsatisfactory" and "advanced") on the four-point observation instrument, which was based on Charlotte Danielson's Framework for Teaching. On any given item, an average of 5 percent of scores were in the bottom category ("unsatisfactory"), while 2 percent of scores were in the top category ("advanced"). The vast majority of scores were in the middle two categories, "basic" and "proficient." (Ho & Kane, 2013)

In any case, effective teacher evaluation relies on the instrument and training processes as much as the observers themselves. One way to align these aspects is to choose an observation instrument that sets clear expectations (Bill and Melinda Gates Foundation, 2012). With an instrument that sets clear expectations, principals can use their contextual knowledge of a teacher in combination with their observations, and outside observers can rely on an outsider perspective.

Use of and Issues with Value-Added Modeling

Another method of teacher effectiveness measurement included in this study is Value-Added Modeling (VAM). In North Carolina, teachers who teach state-tested subjects receive value-added scores through the Educator Value-Added Assessment System (EVAAS®), designed by SAS Institute Inc. and based on the Tennessee Value-Added Assessment System (TVAAS) (Sanders, Saxton, & Horn, 1997). The primary appeal of VAM to measure educator effectiveness is in its ability to capture student growth, rather than simple student achievement.

Value-added models were brought to education with the intent of measuring the specific contribution of a school (or teacher) to a student's learning. Whereas traditional methods of student achievement measurement focus on proficiency and performance level, VAMs involve a prediction model that controls for, at a minimum, a student's prior achievement. When out-of-context student achievement, such as raw achievement scores or the proportion of students above a particular proficiency bar, are used as the primary quantitative basis for teacher evaluation, teachers can be unfairly blamed or helped simply by having a classroom full of students with a history of low or high achievement, respectively. Value-added models (of which there are many forms) use various degrees of student-, classroom-, and school-level covariates (e.g. student demographics, SES, school size) to attempt to isolate the effect of interest. In general, VAMs rely on the assessment of student growth – students are predicted to perform at a particular level on a standardized test determined by the variables in the prediction model, and any deviation from the prediction is attributed to the teacher (and/or the school).

In this sense, VAMs are distinct from more basic measurements of student growth. Growth models, like VAMs, track the same students over time, measuring their performance at

one point in time relative to a previous baseline. However, these simpler growth models implicitly assume that all students learn at a uniform rate and that a teacher (or school) is the only responsible party for growth (Scherrer, 2011). VAMs also track student test score change over time, but in a way that models the added effect of a teacher after controlling for other effects, depending on the VAM used (Timmermans, Doolaards, & de Wolf, 2011).

However, as articulated by Van de Grift, "It is far easier to define the value added of schools than to assess it" (2009, p. 270). Although there are many applications of the value-added concept in educational assessment, one of the more common approaches is a general model from the work of McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004).

Criticisms of VAM. Even though the general aim of VAM is a conceptual improvement over the use of student-based outcome measures without controls as teacher evaluation tools, there remains much criticism about the use and overreliance on VAM. These criticisms exist on both an empirical and a conceptual level, both of which will be addressed in this section.

Van de Grift (2009) enumerates four empirical issues with the use of VAM for teacher evaluation: First, there is typically too much missing data to claim valid teacher effect estimates. Second, he finds that the missing data are not random. Third, VAM results can be highly unstable, and fourth, teacher rankings vary based on thee VAM chosen.

Other studies have criticized VAM for drawing conclusions about teacher or school contributions to student learning based on effect scores that are measuring more than the intended effect. One particularly noteworthy example comes from Rothstein (2010) who, after developing falsification tests for three commonly-used VAMs, was able to show fifth-grade teacher effects as a significant predictor of fourth-grade student score gains, a finding that should be impossible in reality. The author discusses this as an indication that VAM teacher effects contain non-modeled information about the students, such as the non-random sorting of students to schools and teachers, and that these effects should be interpreted with caution.

VAM used in this project. School districts in the state of North Carolina use, as one method of teacher evaluation, the Education Value-Added Assessment System (EVAAS®), developed by SAS Institute Inc. (Wright, White, Sanders, & Rivers, 2010). This particular VAM only controls for a student's prior achievement at the student level and contains no controls at the school level. EVAAS® is based on the TVAAS model (Sanders, Saxton, & Horn, 1997). Despite only controlling for prior achievement, a case can be made that EVAAS® is the most robust and efficient model currently in use due to the model's method of handling missing data and the software's ability to handle large-scale analyses (Amrein-Beardsley, 2008).

Combining Measures of Teacher Effectiveness

Rothstein (2012) pointed out that, whereas much research has been done on the development and validation of teacher effectiveness measures, "relatively little attention has been paid to the design of policies that will use the new measures to improve educational outcomes" (p. 2). Even if there is disagreement on the best methods for assessing teacher quality and the degree to which teacher quality directly impacts student outcomes, evidence suggests that teachers do matter, and that there is variance in teacher quality (Papay, 2012). Sanders (2006) points out that two consecutive years of ineffective teachers can leave a detectable negative impact on future growth

beyond two years of "catch-up" with effective teachers. Hill et al. (2011) noted that teacher impact routinely explains a higher percentage of variance in student achievement than do school- and system-level factors, and other studies have shown that teacher quality strongly impacts such variables as student engagement, student focus, and student performance on other, higher-level assessments (Gersten et al, 2005).

Support for a composite measure. As noted in Stronge et al. (2011), teacher quality is a complex phenomenon, and there is little consensus on what it is or how to measure it. In fact, there is considerable debate as to whether one should judge teacher effectiveness based on teacher inputs, the teaching process, the product of teaching, or a composite of these elements.

Hill et al. (2011) argued strongly in favor of a composite. They emphasized the importance of going beyond simply writing instruments for teacher evaluation and rather creating more in-depth observational systems. These systems would include hiring criteria for raters, rigorous training protocols, robust scoring designs, and quality instruments.

The MET project echoes this view, suggesting that many commonly-accepted measures of teacher effectiveness contribute to fair and accurate assessment of teaching quality. They found that teacher observation scores were strongly related to student achievement in Math and English, and that the relationships between observation scores and student achievement gains were stronger when combined with other information, such as previous student achievement and VAM. Finally, they found that combining measures led to improved reliability and predictive power for future achievement (Bill and Melinda Gates Foundation, 2012, p. 10). Other research provides strong evidence for the use of multiple measures in a teacher evaluation system as well. Jacob and Lefgrin (2008) found that both teacher observations and VAM were statistically significant predictors of future performance.

The use of a combined measure has exhibited content validity as well. Students of teachers scoring high on a combined measure showed larger performance gains on tests of conceptual understanding, higher levels of effort, and greater class enjoyment than students of teachers scoring lower (Bill and Melinda Gates Foundation, 2012, p. 13).

When using multiple measures in a system of evaluation, many factors must be addressed. Darling-Hammond et al. (2012) argued that, in order for a system to be successful, a district must use multiple observations across the year, conducted by expert evaluators looking at multiple sources of data. Staiger & Rockoff (2010) pointed out that high-quality information about teacher effectiveness, taken from multiple sources, can be effective in increasing outcomes through hiring and firing practices, performance-based pay, and targeted professional development. In their "State of the States 2012" (2013) report, the National Teacher Quality Foundation suggested that states use multiple measures of student learning to measure teacher effectiveness, specifically highlighting evidence of student learning observed during classroom observations. Jacob & Lefgrin (2008) pointed out that both principal evaluations of teacher effectiveness and VAM are better indicators of teacher quality than traditional methods of teacher tenure and retention assessment, that is, experience and education level.

With composite measures of teacher effectiveness, teachers who are better at helping students learn can be identified and relatively accurate predictions can be made about teacher performance. If composite measures of teacher effectiveness are attractive, then what aspects of teacher quality should go in to the score? As Teddlie et al. (2006) stated, "having one monolithic dimension, which could only be labeled ''teacher effectiveness,'' certainly does not adequately capture the essence of the construct" (p. 563).

Elements of a composite measure. Johnson (1997) suggested three constructs be considered: the teacher as person, the teaching process, and the teaching product. Johnson also described six key indicators that were common among teachers and could be interpreted as subscales of the three categories: teacher as subject matter expert, teacher as caring, teacher as exhibiting classroom control, teacher as interactive in communication, students as on-task and attentive or engaged, and student progress and achievement.

The MET project suggested combining observation scores and student achievement gains with student feedback. This combination of assessments was found to lead to increased stability in teacher effectiveness scores (Bill and Melinda Gates Foundation, 2013, p. 5). One of the MET Project reports stated that teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students. If a measure is to be included in formal evaluation, they argued, then it should be shown that teachers who perform better on that measure are generally more effective in improving student outcomes (p. 15).

Despite the amount of research appealing for a combined measure including both teacher variables and student outcomes to determine teacher effectiveness, North Carolina (among other states) uses a combination of teacher inputs only to determine pay: teacher experience, advanced degree status (which has since been removed from the 2014 budget), and National Board Certification status. Research has been mixed regarding the impact of these measures on teaching effectiveness.

Effective teaching can be measured, and there are many processes through which to do it. Although much research has been done to assess the mechanics, reliability, and effectiveness of such measures, little research has been done on the most effective way to weight them when establishing a composite score. The aim of this project was to contribute to that discussion by examining and comparing the relationships among three methods of teacher effectiveness measurement.

METHODS

Background

This study took place in a large urban school district in North Carolina. Within this district, 16 schools (12 elementary and 4 middle) have been designated as STAR schools – recipients of a federal Teacher Incentive Fund (TIF) grant. TIF grants are awarded to districts to fund projects with the goal of exploring alternative, merit-based compensation systems for teachers. To be eligible for the grant, districts must base merit pay on teacher evaluation systems that "differentiate effectiveness using multiple rating categories that take into account student achievement growth as a significant factor, as well as classroom observations conducted at least twice during the school year" (Teacher Incentive Fund, 2010, p. 1).

Like all other North Carolina public school teachers, teachers in STAR schools are subject to annual evaluations conducted by their principals. These state-mandated evaluations consist of, among other factors, one to four observations per year using the Standards on the Rubric for Evaluating North Carolina Teachers (McRel, 2009). In addition, many teachers in grades 4-8 in this system receive individual value-added scores through North Carolina's Education Value-Added Assessment System (EVAAS®). Further, teachers in STAR schools are subject to two additional observations per year by trained, full-time observers using a rubric

based on the "Teach" section of Washington, DC Public Schools' IMPACT model. This study is intended to set the stage for further investigation of the use of various teacher evaluation measures (TEMs) in North Carolina and the effects of certain school- and teacher-level variables on classification and score. As the data collected are from a highly specific subset of schools and teachers, the project is framed as a case study and purely descriptive of the particular sample.

Description of Sample

School characteristics. This study examined the relationships among effectiveness measures from teachers at the 16 STAR schools in in the 2011-12 school year. All 16 schools have very high (>80%) FRL percentage, but the schools have a wide range (15.69% - 52.35%) of ELL students. Additionally, all 16 schools have performance composites below the district average.

Instruments

North Carolina Teacher Evaluation System. The NC TES assessment consists of 25 items across five standards. Some of these 25 items are scored based on classroom observations, some are scored based on the collection of artifacts, and some are scored based on a combination of the two. Principals (and in some cases assistant principals) are responsible for the observations, artifact collection, and ultimate evaluation of teachers on the instrument. Each item is scored on a five-point ordinal scale, with labels: (1) Not Demonstrated, (2) Developing, (3) Proficient, (4) Accomplished, and (5) Distinguished. The standards are as follows:

Standard I: Teachers demonstrate leadership (5 items)
Standard II: Teachers establish a respectful environment for a diverse population of students (5 items)
Standard III: Teachers know the content they teach (4 items)
Standard IV: Teachers facilitate learning for their students (8 items)
Standard V: Teachers reflect on their practice (3 items)

In North Carolina, the variability of teacher scores on this instrument for 2011-12 was not large. The vast majority of teachers were rated either "Proficient" or "Accomplished" on all standards (see Table 1). As with the STAR observation tool, there is no published information on the reliability or validity of the NC TES Standards. Each teacher is only observed by one individual, so inter-rater reliability is not available, and raw data on items within standards are not available.

TABLE 1
Proportion of Teachers Receiving Each Rating by Standard on the NC TES, 2011-12

Standard	Not Demonstrated	Developing	Proficient	Accomplished	Distinguished
I	0.1	1.7	36.3	48.4	13.6
II	0.2	2.4	38.7	48.3	10.4
III	0.2	2.6	47.1	41.4	8.7
IV	0.0	2.1	37.6	51.4	8.7
V	0.2	2.6	48.1	39.8	9.3

Source. North Carolina Department of Public Instruction

Depending on the teacher's years of experience, these scores were based on one to four classroom observations per school year and the collection of artifacts. Career status teachers (those with tenure) receive two informal 20-minute observations and one formal 45-minute observation by administrators. Probationary status teachers (those without tenure) receive three formal 45-minute observations and one informal 20-minute observation. In STAR schools, evaluation score data existed for all teachers on standards I and IV, and for 84 teachers on all five standards. This study focused on standards I and IV, as all teachers were assessed on these standards.

EVAAS®. In all NC public schools, teachers who teach subjects for which there is an eligible end-of-year examination (including End-of-Grade, End-of-Course, CTE, and Common Exams), and for whom at least 10 students take the test, receive a value-added score through the Education Value-Added Assessment System (EVAAS®), designed by SAS Institute, Inc.

All teachers who receive EVAAS® scores are evaluated based on their composite scores, regardless of the number of scores comprising the composite. This study focused on the relationship of the teacher evaluation instrument and observation scores with the EVAAS® composite, as all teachers receive composite scores, and since composite scores are more conceptually comparable with the STAR observations and NC TES ratings, which are not subject-specific. In STAR schools in 2011-12, 137 elementary and middle school teachers received EVAAS® composite scores, and these teachers comprise the sample in this study.

STAR observations. Teachers at STAR schools, because of the nature of the grant project, are evaluated by an extra measure of teacher effectiveness. The primary system for teacher evaluation (NC TES) used in the district did not contain an observation protocol that was as detailed as administrators felt was appropriate for the project's goals, so a new observation tool was selected.

After conducting research into observation tools used in other large urban school districts, the "Teach" portion of the Washington, DC Public School System's IMPACT evaluation model was chosen. From a face validity standpoint, the instrument seemed to be broadly applicable, in line with the district's views on effective teaching, and clear in its descriptions. To date, there has been no reliability information published about the tool, nor any published research on the validity of the instrument. In 2011, a conference call between district

officials and DCPS representatives revealed their inter-rater reliability of 0.80 with full-time observers.

In this study, all "core" teachers – defined in STAR as teachers for whom value-added scores are generated – receive two third-party observations per year by trained observers using the "Teach" section of Washington, DC, Public Schools' (DCPS) IMPACT evaluation system. For the purposes of this study, any reference to the STAR observation tool refers to the "Teach" section of the IMPACT rubric used in the STAR project.

The instrument itself features nine standards, with one item per standard. For each standard, the observer assigns a score on a four-point ordinal scale with labels: (1) Ineffective, (2) Minimally Effective, (3) Effective, and (4) Highly Effective. The nine standards are:

Standard 1: Lead Well-Organized, Objective-Driven Lessons
Standard 2: Explain Content Clearly
Standard 3: Engage Students at All Learning Levels in Rigorous Work
Standard 4: Provide Students Multiple Ways to Engage with Content
Standard 5: Check for Student Understanding
Standard 6: Respond to Student Misunderstandings
Standard 7: Develop Higher-Level Understanding through Effective Questioning
Standard 8: Maximize Instructional Time
Standard 9: Build a Supportive, Learning-Focused Community

Teachers were observed by trained observers, all with classroom teaching experience. After a teacher's observation, observers provided both written and verbal feedback to the teacher, and provided copies of the written feedback to the teacher's STAR instructional coach for discussion. Individual teacher scores obtained on STAR observations were not provided to school administrators; only school-level aggregated data were given to administrators.

After inter-rater agreement on full-length lessons was consistently above 0.85 (for dichotomous categorization above/below "minimally effective"), observers conducted live practice observations in pairs in non-STAR schools. Overall inter-rater agreement was found to be 0.91 for dichotomous categorization above/below a score of 2.0), and observers were permitted to begin actual observations in STAR classrooms.

The ultimate goal of this study was to examine the nature of the relationships among principals' evaluations of teachers (that are conducted based on formal and informal observations over the course of a full school year), third-party observations by trained observers without prior relationships with the teachers, and student outcomes. Although the STAR instrument was designed to have nine independent standards, this assumption had not been empirically tested within this district.

An exploratory factor analysis of 2012-13 STAR observation data using principal components analysis led to the conclusion that a one-factor solution was the most reasonable. The one-factor model explained 46% of variance in scores (the second factor added another 10%), and the analysis produced only one eigenvalue with a value greater than one. In addition, the Spearman correlation matrix of the nine standards with N = 866 observations showed significant correlations at the 0.01 level for all 36 pairs of standards ranging from 0.21 to 0.58, with only 5 of 36 correlations below 0.3. The one-factor solution was also shown to be a reliable measure, with an Alpha coefficient of 0.85. Assuming the validity of the one-factor model based on analysis of the 2012-13 data, the 2011-12 STAR observation data was averaged across

standards to produce a single-score indicator to represent quality of teaching as measured by a trained observer.

Data Analysis Procedures

As the focus of this study was primarily a comparison of the classification consistency among the various evaluation methods, the primary method of analysis was cross-tabular analysis. Each teacher was classified into a high-, middle- or low-achievement category on each of the measurements. The null hypothesis that the methods of classification were independent of one another was tested. In all, six tests were conducted – one for each of the six pairs of measurement methods. The NC TES scores are ordinal and categorical in nature, but as seen in Table 1, there were very few teachers classified at level two or below (developing or not demonstrated) on any of the standards. There were also very few teachers classified at level five (distinguished) on any standard. In practice, no reward is given for teachers whose TES scores are in a particular category, but teacher "action plans" are created for a teacher if his or her scores fall into the "developing" category or below.

Like the TES scores, there was no reward for teachers who receive high ratings on the STAR instrument, but there is a potential penalty for very low scores. A teacher who averages below 2.0 on STAR observations receives a 25% reduction in any earned incentive pay. In practice, teachers receiving an average of 3.0 or higher on the STAR instrument are regarded as having taught highly effective lessons. Scores on individual standards on the STAR instrument range from 1-4 (discrete) and the overall score was calculated by taking the mean of the nine standards. For the STAR instrument, scores at or above 1.0 but below 2.0 were classified "low," scores at or above 2.0 but below 3.0 were classified as "middle," and scores at or above 3.0 were considered "high." Mean STAR scores were treated as continuous variables.

In North Carolina, EVAAS® indices divide teachers into three categories – below expected growth (less than -2.0), at expected growth (-2.0 to 2.0),and above expected growth (above 2.0).

RESULTS

To answer the research question, teacher classification via each measurement method was analyzed. First, the distribution of each classification method was explored with descriptive statistics (see Table 2).

TABLE 2
Descriptive Statistics for Each Measurement Method

Method	N	Minimum	Maximum	Mean	S.D.	Skew
EVAAS®	137	-10.11	3.03	-1.13	1.96	-1.02
NC I	137	2	5	3.43	.58	.74
NC IV	137	2	4	3.34	.50	.35
STAR	137	1.63	3.89	2.89	.47	-.52

STAR teacher observation scores are continuous with a possible range of 1.00 – 4.00. NC TES scores are categorical, with a possible minimum of 1 (Not Demonstrated) and a possible maximum of 5 (Distinguished). EVAAS® scores have no theoretical minimum or maximum; they are index scores where scores below zero indicate performance below the state average and scores above zero indicate performance above the state average. Table 3 shows the frequency of scores for each individual measurement.

TABLE 3
Frequency Distributions of NC Standards I and IV for Sample

Standard	Not Demonstrated	Developing	Proficient	Accomplished	Distinguished
NC I	0	1	81	50	5
NC IV	0	2	87	48	0

When examining the EVAAS® data, three cases appeared to be outliers, as they lay below three standard deviations from the mean. The three values, -10.11, -7.66, and -7.21, lay 4.62, 3.36, and 3.13 standard deviations from the mean, respectively. However, the data for the entire district (not just STAR schools) represented a greater range. When the STAR data were examined within the context of the EVAAS® scores for the entire district, these data all fell within 3.6 standard deviations of the mean. Additionally, removal of the three cases would only adjust the sample mean from -1.13 to -.98 – a difference of .15, or .08 of a standard deviation. For these reasons, in addition to having no reason to believe the scores were a result of error, the values were retained.

The scores on the NC TES tended to cluster in categories 3 and 4 (proficient and accomplished), with only a very small proportion of teachers receiving scores outside this range. In North Carolina, teachers are recommended for corrective action if a score of 2 (developing) or below is received. On both standards above, fewer than 1% of teachers received scores lower than 2. If one is to assume that 1% or more of teachers are performing below proficiency in reality, then, the NC TES is not adequately identifying these teachers.

Next, teachers were classified according to each measurement method as outlined previously. Classification criteria for EVAAS® and NC TES are prescribed by the state of North Carolina. Teachers receiving EVAAS® indices below -2.0 are officially considered to be showing less than expected growth. Teachers with EVAAS® indices above 2.0 are officially classified as showing greater than expected growth. On the NC TES, scores of 1 or 2 are indicators of a teacher being "below proficiency," and teachers are subject to personnel action based on this classification. There is no corresponding "high" category, but teachers with a score of 5 are considered "distinguished." In the framework here, scores of 1 or 2 were considered "low," scores of 3 were considered "middle," and scores of 4 or 5 were considered "high." For STAR observations, there is no such statewide classification. In practice in STAR schools, teachers receiving an observation average below 2.0 are subject to a reduction in incentive pay, so the "low" category in this project reflects that classification. Teachers scoring above 3.0 on the four-point scale have scores in the "highly effective" category, and the initial classification of "high" was set at this cut point. Table 4 shows the distribution of teachers in each classification by each measurement method.

TABLE 4
Classification Distributions by Measurement Method

	Classification		
Method	Low	Middle	High
EVAAS®	28%	68%	4%
NC I	1%	59%	40%
NC IV	1%	64%	35%
STAR	4%	48%	48%

Except for EVAAS®, the measures rarely placed teachers in the "low" category. The STAR classification included very few teachers in the "low" classification, and relatively even amounts in the "middle" and "high" categories. For NC Standards I and IV, almost no teachers were classified as "low," the majority of teachers were placed in the "middle," and 35-40% placed in the "high" category.

To determine if there is a statistically significant difference in classification, a series of cross tabular analyses were conducted (see Table 5), with the intention of the Pearson Chi-square being the statistic of interest.

TABLE 5
Cross-tabulation of Teachers by Effectiveness Category and Measurement Method

	EVAAS			STAR			NC I			
	Low	Middle	High	Low	Middle	High	Low	Middle	High	Total
STAR										
Low	5	0	0							5
Middle	23	43	0							66
High	10	50	6							66
NC I										
Low	0	1	0	0	1	0				1
Middle	29	49	3	4	44	33				81
High	9	43	3	1	21	33				55
NC IV										
Low	1	1	0	0	2	0	0	2	0	2
Middle	32	52	3	5	47	35	1	67	19	87
High	5	40	3	0	17	31	0	12	36	48
Total	38	93	6	5	66	66	1	81	55	

For each of these tests, the null hypothesis being tested was that there is no association between the two methods of classification. However, due to the low number of teachers categorized as "low" by both STAR and the NC TES measures and the low number of teachers categorized as "high" by EVAAS®, all tables contain 5/9 (55.6%) cells with an expected count below 5. Chi-square analyses were not considered appropriate in such conditions. Therefore, Cohen's Kappa was calculated for each analysis as a measure of agreement (see Table 17).

With four methods of classification (EVAAS®, STAR, NC I, and NC IV), there were six pairs of methods to compare. With repeated statistical tests, it is wise to correct for a capitalization on chance when determining the rejection level (alpha) of a test statistic. For this project, a Bonferroni correction was applied. An overall alpha of .05 was desired, so with six comparisons, the alpha for each individual test was set at .008.

TABLE 6
Agreement of Measurement Methods

Test	Cohen's Kappa	p
STAR - EVAAS®	0.06	0.17
NC I - EVAAS®	-0.07	0.09
NC IV - EVAAS®	-0.08	0.08
NC I - STAR	0.16	0.04
NC IV - STAR	0.18	0.02
NC I - NC IV	0.49	<0.001

As can be seen in Table 6, no pair of measurement methods exhibited statistically significant agreement, with the exception of the NC I – NC IV pair. Although the NC I – STAR and NC IV – STAR pairs exhibited a p-value below .05, a Bonferroni correction for six hypothesis tests yields a critical alpha of .008. From a policy perspective, substantial disagreement between methods of teacher effectiveness measurement could be problematic, particularly if there is a pattern of teachers scoring high on one measure while scoring low on another. The following are the measurement disagreements shown in Table 5 where these "major" disagreements (disagreement by more than one category) occurred:

1. A teacher categorized as "low" by EVAAS® (n = 38) was twice as often categorized as "high" by STAR (n = 10) than as "low" (n = 5).
2. A teacher categorized as "high" by STAR (n = 66) was almost twice often categorized as "low" by EVAAS® (n = 10) than "high" (n = 6).
3. A teacher categorized as "high" by NC I (n =55) was three times as often categorized as "low" by EVAAS® (n = 9) than "high" (n = 3).
4. A teacher categorized as "low" by EVAAS® (n = 38) was more often categorized as "high" by NC IV (n = 5) than "low" (n = 1).
5. A teacher categorized as "high" by NC IV (n = 48) was more often categorized as "low" by EVAAS (n = 5) than "high" (n = 3).

These findings only address the disagreement between measurement methods when disagreement was by at least two categories (that is, when teachers were classified as both "low" and "high" by different measures). These types of major disagreements are the most problematic from a policy perspective.

As seen in Table 4, 48% or more of teachers are categorized into the "middle" by each measurement method. This high proportion of teachers being into the "middle" by all methods causes the highest marginal likelihood of any teacher, given any classification by any method, to be "middle," with the following exceptions:

1. Teachers rated “low” by STAR were most often rated “low” by EVAAS.
2. Teachers rated “middle” by EVAAS were most often rated “high” by STAR.
3. Teachers rated “high” by EVAAS were most often rated “high” by STAR.
4. Teachers rated “high” by EVAAS were equally often rated “middle” or “high” by NC I.
5. Teachers rated “high” by EVAAS were equally often rated “middle” or “high” by NC IV.
6. Teachers rated “high” by STAR observation were equally often rated “middle” or “high” by NC I.
7. Teachers rated “high” by NC I were most often rated “high” by STAR observation.
8. Teachers rated “high” by NC IV were most often rated “high” by STAR observation.
9. Teachers rated “high” by NC IV were most often rated “high” by NC I
10. Teachers rated “high” by NC I were most often rated “high” by NC IV.

Given the analysis of classification agreement, the answer to the research question – whether the measurement methods classify teachers in substantively different ways – would be yes. There was substantive disagreement between methods in the classification of teachers.

DISCUSSION

This project examined teacher effectiveness data from 16 schools in a large urban district in North Carolina. The 16 schools were all participants in a Federal Teacher Incentive Fund grant program, and as such, were all high-poverty, low-achieving schools. Data collected from these teachers included a value-added measure of student growth, a classroom observation measure conducted by trained observers with a standardized rubric, and two measures required by the state to be scored by principals. Of the two principal ratings, one was a rating of teacher leadership and the other was a measure of pedagogy.

With much research currently being done on the effective use of teacher evaluation measures and their appropriate place in hiring, retaining, promoting, and firing teachers, this project aimed to contribute to the discussion by performing in-depth descriptive analyses on the effects of four distinct teacher evaluation measures. Of particular interest was the sorting of teachers into performance classes by each instrument and the variables that affect such groupings. In these specific schools, the principal-rated measures are potentially used to make personnel actions at the low end of the scale, but are not used in determining any positive interventions such as teacher pay. Conversely, the value-added measure of student growth is used to reward high-achieving teachers with incentive pay, while the classroom observation measure is used to reduce the incentive pay, if classification differences exist where the teacher earns high value-added scores but low observation scores. Observation results are also used as formative assessments for teacher development and growth.

The research question addressed in the project was, “Do the methods of teacher effectiveness measurement classify teachers in substantively different ways?” To answer this question, distributions were created and each teacher was classified into a “high,” “middle,” or “low” performance group by each measurement method. Cross tabulations were analyzed on each pair of measurement methods to determine if the methods classified teachers in meaningfully different ways.

When examining the classification of teachers on the value-added and observation methods more closely, every teacher classified as "high" by EVAAS® was also classified as such by STAR, and every teacher classified as "low" by STAR was also classified as such by EVAAS®. However, there was a large amount of disagreement in classification between the two measures. For example, teachers classified as "low" by EVAAS® were twice as often categorized as "high" by STAR than as "low" by STAR. Similarly, teachers categorized as "high" by STAR were almost twice as often categorized as "low" by EVAAS® than "high".

Among teachers categorized as "high" by NC I, three times as many teachers were categorized as "low" by EVAAS® than "high" by EVAAS®. Among teachers categorized as "low" by EVAAS®, more teachers were categorized as "high" by NC IV than as "low" by NC IV. Among teachers categorized as "high" by NC IV, teachers were more often categorized as "low" by EVAAS® than as "high" by EVAAS®. In essence, EVAAS® appears to have categorized teachers very differently than the other methods.

The most meaningful finding was that only the two principal measures (NC I and NC IV) were shown to moderately agree on classification. The fact that these measures of teacher effectiveness that collect vastly different types of data tended to be positively associated before classification speaks to the existence of a general teacher effectiveness value, and that the instruments were measuring it to some degree, although agreement among classifications does not reflect this. This finding also resembles that of Gersten et al. (2005) who found moderate correlations between observation data and student growth data for teachers.

However, there was much more disagreement than agreement when analyzing the categorization of teachers across the different instruments. In particular, the amount of disagreement across two categories (such as when a teacher is categorized as "high" on one measure and "low" on another, rather than "high" and "middle") is a disturbing finding if these results are to be used interchangeably. That is, if a district or state chooses to use only some of these four methods, valuable information about teacher quality may be left out. A teacher who is categorized as "high" by EVAAS® and "middle" by NC I or IV may very well be categorized as "low" by the STAR instrument, but the STAR instrument is not used in North Carolina schools outside of the STAR project. In fact, in this sample, teachers categorized as "high" by EVAAS® are more often rated "low" than "high" on STAR. This suggests that either the two instruments are capturing very different dimensions of teaching quality, that one (or both) are measuring something unrelated to teacher quality, or that the methods are unreliable. Without evidence to suggest that the STAR observation is unreliable or invalid, adding the information provided by the STAR instrument actually provides a more complete picture of the teacher's quality than EVAAS® and the NC TES alone. The STAR tool also has the added benefit of providing formative information to teachers, where EVAAS® does not (at least in terms of pedagogical practice).

The argument for adding STAR observation scores, or some other appropriate third-party observation measure to the measures currently in use in NC is stronger when one considers that formative feedback is a part of the STAR observation. In this sense, information is provided to help administrators identify effective and ineffective teachers, but teachers also benefit by gaining another, detailed perspective into the quality of their practice. However, as seen in the results, STAR score classification doesn't tend to agree highly with the other methods measured. A composite measurement system consisting of individual measures that disagree more often than they agree strains the credibility of the entire system. If none of these measures is universally recognized as the measure that most accurately captures what it means to be an

effective teacher, then issues arise when associating these scores with rewards or personnel actions. Further investigation into the cause of this disagreement is recommended. The inter-rater reliability of the STAR observation tool has been shown to be high. The NC TES is the only measurement method analyzed in this project that has no published reliability or validity. With such high disagreement between measurement methods in this project, an investigation into the reliability of the NC TES might yield helpful results, particularly as the full NC TES currently comprises five of the six aspects of a teacher's official state evaluation.

In summary, the data analyzed in this project have shown that principal evaluations and STAR observations tend to classify teachers as moderate and effective much more often than ineffective, and that EVAAS® scores tend to classify teachers as ineffective and moderate much more often than effective (see Table 4). In general, all methods studied tend to disagree on teacher classification. The use of the STAR observation scores seems to be a valid addition to the evaluation system applied in STAR project schools, as it provides another level of usable data with regard to identifying effective teachers. The STAR observations serve the added benefit of providing formative feedback for teachers who are concerned with professional growth – an element that is otherwise lacking. The inclusion of a standards-based observation measure such as STAR would be crucial if the importance of using EVAAS® to determine personnel actions grows, since not all teachers who are exhibiting highly effective pedagogy are showing high levels of student growth. The two measures are related, but not redundant. The incorporation of a third-party standards-based observation component to the state teacher evaluation system would also follow recommendations discussed by Darling-Hammond et al. (2012), Jacob & Lefgrin (2008), Staiger & Rockoff (2010), and Teddlie (2006).

As mentioned throughout the study, the sample size for this project was quite small and highly specific. All teachers sampled worked in high-poverty schools that had been identified as high-need for the purposes of a federal grant intervention. It is very possible that findings from this study would not translate to a more diverse set of schools and teachers. Not only did the small sample size affect the data itself, in that lower-income schools and lower-growth student bodies were overrepresented relative to the state, but also in the statistical power of most analyses. Findings that may be true in the population would be difficult to perceive without a larger sample size. In addition, much of the data used violates assumptions of normality, further harming the generalizability of findings.

Another limitation of a study of this nature speaks to the difficulty in measuring teacher effectiveness at all. That is, "truth" is unknown. Comparing multiple measures of a single construct is difficult when none of the measures have been satisfactorily validated. When it is shown that STAR is more likely to identify teachers as highly effective than EVAAS®, for example, any conclusions are based purely on definition by the other instrument – hardly an ideal condition.

This particular project, because of its limitations, provides multiple directions for future research. The first and most obvious direction for a future study would be to replicate the analysis with a larger, more diverse sample – perhaps an entire district or state. Some of the findings of this project would carry a strong argument for policy revision if they were found in a sample more representative of the state at large. The structure of the study would remain intact. A larger number of schools randomly selected from the entire state of North Carolina could be sampled, and the same data – EVAAS® indices, principal evaluations, and third-party standards-based observation scores (if available) could be examined. A larger, more representative sample

would provide more diversity in growth scores and demographic variables, and more statistical power for analyses.

With the availability of data on teacher effectiveness growing annually, the emphasis on identifying high-quality teachers will continue to grow. Ensuring that teachers have quality feedback and direction toward this end is crucial in increasing student learning. Stringent analysis of the instruments and methods used in these processes will serve to ensure that teacher evaluation remains fair, efficient, and beneficial.

REFERENCES

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

Anderson, J. (2013). Curious grade for teachers: Nearly all pass. *The New York Times*. Retrieved from http://www.nytimes.com/2013/03/31/education/curious-grade-for-teachers-nearly-all-pass.html?pagewanted=all.

Arrington, S. E. (2010). *Elementary principals' follow-through in teacher evaluation to improve instruction.* (Doctoral Dissertation). Georgia Southern University, Statesboro, GA.

Banchero, S. & Kesmodo, D. (2011). Teachers are put to the test. *The Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424053111903895904576544523666

Bill and Melinda Gates Foundation. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Policy and Practice Summary. *MET Project.*

Bill and Melinda Gates Foundation. (2013). Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study. Policy and Practice Brief. *MET Project.*

Collins, C., and Amrein-Beardsley, A. (2013). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116(1).*

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

Gersten, R., Baker, S. K., Haager, D., & Graves, A. W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial & Special Education, 2(4),* 197-206.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48(3),* 794-831.

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

Ingle, K., Rutledge, S., & Bishop, J. (2011). Context matters: principals' sensemaking of teacher hiring and on-the-job performance. *Journal of Educational Administration, 49(5),* 579-610.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26(1),* 101-136.

Johnson, B. L. J. (1997). An organizational analysis of multiple perspectives of effective teaching: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11(1),* 69-87.

Johnson, S. D., & Roellke, C. F. (1999). Secondary teachers' and undergraduate education faculty members' perceptions of teaching-effectiveness criteria: A national survey. *Communication Education, 48(2),* 127-138.

Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, *86(10),* 727-735.

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.

McConney, A., Ayres, R., Hansen, J. B., & Cuthbertson, L. (2003). Quest for quality: recruitment, retention, professional development, and performance evaluation of teachers and principals in Baltimore City's public schools. *Journal of Education for Students Placed at Risk, 8(1),* 87-116.

McREL. (2009). *North Carolina Teacher Evaluation Process*. Teacher Evaluation Instrument. Retrieved from http://www.ncpublicschools.org/docs/effectiveness-model/ncees/instruments/teach-eval-manual.pdf

(MET) Measures of Effective Teaching Project - K-12 Education. (n.d.). Retrieved from http://k12education.gatesfoundation.org/teacher-supports/teacher-development/measuring-effective-teaching/

North Carolina Department of Public Instruction – Educator Effectiveness Data. (n.d.). Retrieved from http://apps.schools.nc.gov/pls/apex/f?p=155:5:1500221441455501::::P5_YEAR:2011-12%20School%20Year

National Council on Teacher Quality, (2013). *State of the states 2012: Teacher effectiveness policies*. Washington, DC:

Noakes, L. A. (2008). Adapting the utilization-focused approach for teacher evaluation. *Journal of Multidisciplinary Evaluation, 6(11),* 83-88.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82(1),* 123-141.

Ramirez, A., Lamphere, M., Smith, J., Brown, S., & Pierceall-Herman, J. (2011). Teacher development and evaluation: A study of policy and practice in Colorado. *Management in Education, 25*(3), 95-99.

Ripley, A. (2010, January 1). What Makes a Great Teacher? *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/2010/01/what-makes-a-great-teacher/307841/

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125(1),* 175-214.

Rothstein, J. (2012). *Teacher quality policy when supply matters*. (No. w18419). National Bureau of Economic Research.

Sanders, W. (2006, October). Comparisons among various educational assessment value-added models. In *National Conference on Value-Added*.

Sanders, W., Saxton, A. & Horn, S. (1997). The Tennessee value-added assessment system. In Millman, J. (Ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95(2),* 122-140.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives, 24*(*3),* 97-117.

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.

Teacher Incentive Fund. (2010). Summary of notice of proposed priorities. Retrieved from http://www2.ed.gov/programs/teacherincentive/summarymemo32010.doc

Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., & Yu, F. (2006). The international system for teacher observation and feedback: Evolution of an international study of teacher effectiveness constructs. *Educational Research and Evaluation, 12(6),* 561-582.

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement, 22(4),* 393-413.

Toch, T. (2008). Fixing teacher evaluation. *Educational Leadership, 66(2),* 32-37.

Van de Grift, W. (2009). Reliability and validity in measuring the value added of schools. *School effectiveness and school improvement, 20(2),* 269-285.

Van Tassel-Baska, J., Quek, C., & Feng, A. X. (2006). The development and use of a structured teacher observation scale to assess differentiated best practice. *Roeper Review, 29(2),* 84-92.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS® EVAAS® statistical models. *SAS White Paper*.

Zatynski, M. (2012). Revamping teacher evaluation. *Principal, 91(5),* 22-27.

Zimmerman, S., & Deckert-Pelton, M. (2003). Evaluating the evaluators: Teachers' perceptions of the principal's role in professional evaluation. *NASSP Bulletin, 87(636),* 28-37.