MARZANO
Evaluation Center

Evidence for the Marzano Focused Model: Reforming Teacher Evaluation in a Time of Change

REPORT

Lindsey Devers Basileo
Michael D. Toth

2020

1-866-731-1999 | MarzanoEvaluationCenter.com
Marzano Evaluation Center is a division of Instructional Empowerment, Inc.

Table of Contents

2 Abstract
3 Introduction
5 Literature Review
  5 Constraints for School Leaders in Evaluation
  6 The Marzano Focused Teacher Evaluation Model
  8 Evidence for the Marzano Teacher Evaluation Model
9 Methods
  9 Outcome Variable: Teacher Value-Added Measures
  10 Predictor Variables
12 Findings
16 Conclusion
17 Policy Implications
18 References

Abstract

The Every Student Succeeds Act (ESSA) calls for fairer teacher evaluation and support systems that are based on evidence of student achievement. The law calls for developing and disseminating high-quality evaluation tools, such as classroom observation rubrics. In response to this legislation, the Marzano Teacher Evaluation Model (MTEM) was redesigned to be more succinct and to have a greater focus on improving teacher performance. The newly designed model, termed the Marzano Focused Teacher Evaluation Model (FTEM) aims to be less time consuming, more accurate, and more fair. The current study will investigate the relationship between teacher observation ratings and value-added measures, examining correlation levels of both models. The study extracted observation ratings from a collection platform and linked it to teacher value-added measures in Florida. The major finding from this study is that regardless of whether the classic MTEM or new FTEM is employed, the magnitude of correlations are surpassed. Additionally, the random intercept models showed that the observation score was the largest and statistically significant predictor of teacher value-added measures controlling for student, observation system characteristics and school poverty. While more research is necessary, the redesigned model meets the needs of the ESSA legislation.

Introduction

Closing student achievement gaps remains a national priority as demonstrated by the United States' national education law, Every Student Succeeds Act (ESSA). The ESSA was passed in December 2015, and while it modified its predecessor the No Child Left Behind Act (NCLB, 2002), it did not eliminate provisions relating to standardized assessments. Like the No Child Left Behind Act, the ESSA is a reauthorization of the 1965 Elementary and Secondary Education Act, which established and expanded the federal government's role in public education. The legislation builds upon the previous reauthorization to identify key areas of progress and improve educational experiences for all students. While the NCLB included legislative reforms that focused on high stakes evaluative measurement systems for teachers and principals (Alger, 2012; Auguste, Kihn, & Miller, 2010; Bill & Melinda Gates Foundation, 2012; Renter, 2012), the ESSA has provided states more discretion in measuring student, teacher, and school performance. States can now decide to what extent they weigh students' test scores as components of their revised teacher evaluation policies and systems.

The ESSA legislation calls for fairer teacher evaluation and support systems that are based on evidence of student achievement (ESSA, §2103). The law calls for developing and disseminating high-quality evaluation tools, such as classroom observation rubrics, to incorporate evaluation results to inform decision-making about professional development and improvement strategies. Teacher evaluation systems are critical to these movements as they are the formal process a school uses to review and rate teachers' performance and effectiveness (ESSA, §2101). It is through evaluation systems where teachers receive feedback to improve instruction, with the goal of increasing teacher pedagogy and increasing student achievement. While there has been a slight shift in terms of the extent students' test scores will be weighted, increasing student achievement remains a national priority. Consequently, incorporating teacher evaluation systems which are predictive of student achievement remain a critical policy nationwide.

In response to this legislation, in addition to the demand from school leaders to have a more efficient evaluation model, the Marzano Teacher Evaluation Model (MTEM) was revised to be more succinct, measurable, and to have a greater focus on improving teacher performance. The redesigned model, termed the Marzano Focused Teacher Evaluation Model (FTEM), was launched in the 2017–2018 school year and several districts in Florida adopted the framework. This provided a prime opportunity to assess the predictability of the new model as both the classic and updated models were utilized in the field during the same timeframe. The MTEM is widely used in

Florida and previous studies have validated the model in terms of predictability of student growth (Basileo & Toth, 2019). Whether the FTEM can uphold the same magnitude of effects as the classic model is the focus of this study. The current study will also outline the changes of the reformed model and provide evidence of predictability. Evidence-based teacher evaluation systems are vital to meeting the national priorities set forth in the ESSA, particularly when teacher evaluation and observation scores are predictive of student achievement.

The current study will examine the relationship between teacher observation ratings and teacher value-added measures, exploring the correlation levels of both the classic model condensed into the new competencies and the reformed model in the state of Florida for the 2017–2018 school year. Hierarchical linear modeling is used to test whether observation scores predict teacher value-added measures. Value-added measures, sometimes also referred to as growth measures, are used to estimate how much impact teachers have on student achievement during a school year. Value-added models isolate a teacher's contribution by controlling for student, classroom, and school-level measures, making it possible to study individual growth.

Literature Review

Constraints for School Leaders in Evaluation

Principals with strong instructional leadership have been shown to have positive effects on student achievement (Robinson, Lloyd, & Rowe, 2008). Principals are key to successful teacher practice, as their influence on effective teaching is founded on a teacher evaluation system that focuses on providing meaningful feedback, mentoring, and coaching to teachers. Unfortunately, time is an overwhelming concern for school leaders, particularly as districts continue to eliminate support positions (Childress, 2014). Because time is such a scarce commodity for principals, districts are looking for evaluation systems that are efficient and time effective.

A 2013 survey of members included in the National Association of Elementary School Principals (NAESP) and the National Association of Secondary School Principals (NASSP) found that teacher evaluation requires an average of about 13 hours per teacher over the course of a school year (Childress, 2014). Furthermore, principals manage an average staff of 25 in small schools and 60 in large schools. This equates to about 8 and 20 working weeks principals should be dedicating to teacher evaluation within one school year. Additionally, instructional coaching accounts for only 12.7% of principals' time (Grissom, Loeb, & Master, 2013). The largest chunk of that time was spent on informal classroom walkthroughs (5.7%) and only 1.8% of time was spent on formally evaluating teachers. Thus, the ability to provide meaningful and actionable feedback for every teacher poses a major challenge to school leaders given their full range of responsibilities.

While teacher evaluation makes up a small part of a principal's overall responsibilities, accuracy in scoring is essential during formal observations – particularly in high-stakes systems and for reliability and model predictability.

While teacher evaluation makes up a small part of a principal's overall responsibilities, accuracy in scoring is essential during formal observations – particularly in high-stakes systems and for reliability and model predictability. The number of lessons observed for each teacher, and the number of different observers or raters, can vary greatly by school and district. This variation impacts the reliability of the estimates even when there is reliability between raters (Hill, Charalambouse, & Kraft, 2012). Teachers who have more raters and more scores should theoretically have higher teacher value-added

scores because they have received more feedback. Change is imperative for observational frameworks to be less time consuming and more accurate, particularly with the ESSA legislation calling for additional fair evaluation systems that are based on evidence of student achievement and that results in informed decision-making about improvement strategies.

The Marzano Focused Teacher Evaluation Model

Feedback from the field echoed similar concerns found in the ESSA legislation. School leaders needed a more concise evaluation model that supports inter-rater reliability, more accurately measures teacher practice, and increases efficiency to improve performance all while reducing the amount of time spent on observations. The classic model draws from research articulated in Robert Marzano's *The Art and Science of Teaching* (2007) and from earlier works, including: *What Works in Schools* (Marzano, 2003), *Classroom Instruction that Works* (Marzano, Pickering, & Pollock, 2001), *Classroom Management that Works* (Marzano, Marzano, & Pickering, 2003), and *Classroom Assessment and Grading that Work* (Marzano, 2006). The classic model incorporates 60 research-based elements collapsed into four domains. The model was redesigned to be more succinct, measurable and to have a greater focus on improving teacher performance. The FTEM incorporates all the same concepts as the classic model; however, it has collapsed the 60 elements into 23 core competencies while maintaining the four domains. Figure 1 illustrates the classic model redesigned into the FTEM. The classic model is represented within the domains on the outside of the figure while the new model is shown in the center.

School leaders needed a more concise evaluation model that supports inter-rater reliability, more accurately measures teacher practice, and increases efficiency to improve performance all while reducing the amount of time spent on observations.

The new model aims to return time back to administrators so they can focus on instructional coaching, advancing their professional development, and providing feedback to teachers. The FTEM is more streamlined, and it has less competencies to focus on when conducting observations, thereby reducing time and improving the accuracy of scoring. A more focused model, in turn, helps administrators provide teachers with more concise feedback, emphasizing fewer areas to learn while strengthening professional growth.

An additional change to the newly designed model included recommending a competency-based scoring approach to support teacher growth and improve fairness. In other words, by the end of the school year, teachers should have a score on each of the competencies in the model (Carbaugh, Marzano, & Toth, 2017).

Domain 2: Planning and Preparing

Planning and Preparing

Planning and Preparing for Lessons and Units
42. Effective Scaffolding of Information within Lessons
43. Lessons within Units
44. Attention to Established Content Standards

Planning and Preparing for Use of Resources andTechnology
45. Use of Available Traditional Resources
46. Use of Available Technology

Planning and Preparing for the Needs of English Language Learners
47. Needs of English Language Learners

Planning and Preparing for the Needs of Students Receiving Special Education
48. Needs of Students Receiving Special Education

Planning and Preparing for the Needs of Students Who Lack Support for Schooling
49. Needs of Students Who Lack Support for Schooling

Domain 1: Classroom Strategies and Behaviors

Lesson Segment Enacted on the Spot

DQ5: Engaging Students
24. Noticing When Students are Not Engaged
25. Using Academic Games 26. Managing Response Rates
27. Using Physical Movement
28. Maintaining a Lively Pace
29. Demonstrating Intensity and Enthusiasm
30. Using Friendly Controversy
31. Providing Opportunities for Students to Talk about Themselves
32. Presenting Unusual or Intriguing Information

DQ7: Recognizing Adherence to Rules and Procedures
33. Demonstrating "Withitness"
34. Applying Consequences for Lack of Adherence to Rules and Procedures
35. Acknowledging Adherence to Rules and Procedures

DQ8: Establishing and Maintaining Effective Relationships with Students
36. Understanding Students' Interests and Backgrounds
37. Using Verbal and Nonverbal Behaviors that Indicate Affection for Students
38. Displaying Objectivity and Control

DQ9: Communicating High Expectations for All Students
39. Demonstrating Value and Respect for Low Expectancy Students
40. Asking Questions of Low Expectancy Students
41. Probing Incorrect Answers with Low Expectancy Students

Domain 1: Classroom Strategies and Behaviors

Lesson Segment Addressing Content

DQ2: Helping Students Interact with New Knowledge
6. Identifying Critical Content
7. Organizing Students to Interact with New Content
8. Previewing New Content
9. Chunking Content into "Digestible Bites"
10. Helping Students Process New Content
11. Helping Students Elaborate on New Content
12. Helping Students Record and Represent Knowledge
13 Helping Students Reflect on Learning

DQ3: Helping Students Practice and Deepen New Knowledge
14. Reviewing Content
15. Organizing Students to Practice and Deepen Knowledge
16. Using Homework
17. Helping Students Examine Similarities and Differences
18. Helping Students Examine Their Reasoning
19. Helping Students Practice Skills, Strategies, and Processes
20. Helping Students Revise Knowledge

DQ4: Helping Students Generate and Test Hypotheses
21. Organizing Students for Cognitively Complex Tasks
22. Engaging Students in Cognitively Complex Tasks Involving Hypothesis Generation and Testing
23. Providing Resources and Guidance for Cognitively Complex Tasks

Standards-Based Planning

- Planning Standards-Based Lessons/Units
- Aligning Resources to Standard(s)
- Planning to Close the Achievement Gap Using Data

Standards-Based Instruction

- **Identifying Critical Content from the Standards**
- Previewing New Content
- Helping Students Process New Content
- Using Questions to Help Students Elaborate on Content
- Reviewing Content
- Helping Students Practice Skills, Strategies, and Processes
- Helping Students Examine Similarities and Differences
- Helping Students Examine Their Reasoning
- Helping Students Revise Knowledge
- Helping Students Engage in Cognitively Complex Tasks

Conditions for Learning

- Using Formative Assessment to Track Progress
- Providing Feedback and Celebrating Progress
- Organizing Students to Interact with Content
- Establishing and Acknowledging Adherence to Rules and Procedures
- Using Engagement Strategies
- Establishing and Maintaining Effective Relationships in a Student-Centered Classroom
- Communicating High Expectations for Each Student to Close the Achievement Gap

Professional Responsibilities

- Adhering to School and District Policies and Procedures
- Maintaining Expertise in Content and Pedagogy
- Promoting Teacher Leadership and Collaboration

Marzano Focused Teacher Evaluation Model

Domain 1: Classroom Strategies and Behaviors

Lesson Segment Involving Routine Events

DQ1: Communicating Learning Goals and Feedback
1. Providing Rigorous Learning Goals and Performance Scales (Rubrics)
2. Tracking Student Progress
3. Celebrating Success

DQ6: Establishing Rules and Proceduresk
4. Establishing Classroom Routines
5. Organizing the Physical Layout of the Classroom

Domain 3: Reflecting on Teaching

Reflecting on Teaching

Evaluating Personal Performance
50. Identifying Areas of Pedagogical Strength and Weakness
51. Evaluating the Effectiveness of Individual Lessons and Units
52. Evaluating the Effectiveness of Specific Pedagogical Strategies and Behaviors

Developing and Implementing a Professional Growth Plan
53. Developing a Written Growth and Development Plan
54. Monitoring Progress Relative to the Professional Growth and Development Plan Pedagogical Strategies and Behaviors

Domain 4: Collegiality and Professionalism

Collegiality and Professionalism

Promoting a Positive Environment
55. Promoting Positive Interactions with Colleagues
56. Promoting Positive Interactions about Students and Parents

Promoting Exchange of Ideas and Strategies
57. Seeking Mentorship for Areas of Need or Interest
58. Mentoring Other Teachers and Sharing Ideas and Strategies

Promoting District and School Development
59. Adhering to District and School Rules and Procedures
60. Participating in District and School Initiatives

Figure 1. The Transformation to the Marzano Focused Teacher Evaluation Model

For the classic model, raters were directed to score the most dominant elements observed (Basileo, 2016). This new recommendation is a fairer system of scoring, as all teachers have scores on each competency. Consequently, the shift to fewer competencies decreases the amount of time it takes to learn, observe, and rate observations, providing an overall more accurate scoring system with more focused feedback. While the theoretical arguments for modifying the framework are strong and practical, this study will directly test how the reformed model predicts student achievement. Whether the FTEM takes less time to implement falls out of the realm of this study, but should be a focus of future research. Prior to detailing the study methodology, the literature on the classic model will be outlined, focusing on the established relationship between observation scores and teacher value-added measures.

Evidence for the Marzano Teacher Evaluation Model

The largest empirical investigation to assess the predictability of the Marzano Teacher Evaluation Model (MTEM) was conducted by Basileo and Toth (2019). First, they thoroughly reviewed the extant literature from all teacher evaluation frameworks, focusing on studies that have investigated the correlations between teacher observation ratings and teacher value-added measures. From their review, they found small to moderate correlations between observation scores and teacher value-added measures regardless of the rubric utilized to obtain the observation score. Almost all associations were positive. 64% of the coefficients were statistically significant in English Language Arts (ELA) and 43% were statistically significant in math. They found similar correlation levels in their own investigation which included observation ratings and value-added measures of over 12,000 teachers collected in Florida using three years of data (2012–2013, 2013–2014, and 2014–2015). Additionally, the authors used multilevel modeling and found that the observation score was the largest, statistically significant predictor in the model accounting for student, teacher, observation system, and school-level characteristics.

There are several other small-scale studies that have investigated the validity of the MTEM. Haystead and Marzano (2009) synthesized approximately 300 small-scale studies conducted at the Marzano Research Laboratory, and indicated that on average, the elements within the model were associated with an effect size of .42. Another study conducted by the Marzano Research Laboratory (2011) investigated correlations using observation data from 19 to 54 teachers. The study found small to moderate correlations across the 41 elements in domain one. Lastly, Alexander (2016) investigated the relationship between the school value-added measure and average teacher evaluation ratings in 29 districts implementing the MTEM in Florida. Across three years of data, they found small and statistically significant correlations across the schools. Because the FTEM is a newly designed model, there is no empirical evidence on the model to date. This study will be the first to assess the impacts for the reformed evaluation framework.

Methods

The current study will investigate the relationship between teacher observation ratings and teacher value-added measures using correlation levels for the condensed version of the MTEM and the FTEM. Basileo and Toth (2019) have detailed the correlations across three years of data for the classic model. This study will extend beyond theirs by using observation scores collected from the reformed model and by collapsing the elements within the MTEM to create average competency scores for the FTEM. Recall that the FTEM reduced the number of competencies; however, the concepts of the original 60 elements were condensed into 23 competencies as noted in Figure 1. Thus, two datasets will be investigated: the first dataset examines correlation levels of a condensed version of the MTEM for those districts who were implementing the classic model, and the second assesses correlation levels for those districts that implemented the FTEM during that same school year. Additionally, hierarchical linear modeling will be used to test whether observation scores predict teacher value-added measures, controlling for observation system characteristics and poverty within the area the school is located.

Outcome Variable: Teacher Value-Added Measures

The dependent variables were obtained from the Florida Department of Education (FLDOE). The secondary dataset included aggregated teacher value-added measures in ELA, math, and a combined (ELA and math) value-added measure for the 2017–2018 school year. The FLDOE value-added model estimates the effectiveness of a teacher by isolating the contribution of the teacher to student learning. Predicted scores are based on prior testing history and student-level characteristics, compared to how well other students in the state perform in that same grade level.

The value-added modeling techniques implemented in Florida are covariate adjustment models that include up to two prior assessment scores and student-level characteristics, including:

- Prior achievement measure(s),
- The number of subject-relevant courses,
- English Learner (EL) status,
- Gifted status,
- Student attendance,
- Student mobility,
- Difference in modal age of the grade level,
- Class size, and
- Homogeneity of students' entering test scores.

These variables are incorporated in the FLDOE model to isolate differences in teachers' classrooms. Below is the general equation used to create the teacher value-added measure and an excerpt from the Florida Value-Added Model Technical Report that explains the equation in the subsequent footnote (2013, pp. 6). [1]

[1] y_{ti} is the observed score at time t for student i, $\mathbf{X}_i$ is the model matrix for the student and school level demographic variables, $\boldsymbol{\beta}$ is a vector of coefficients capturing the effect of any demographics included in the model, $y_{t-r,i}$ is the observed lag score at time t-r ($r \in \{1,2,\ldots,L\}$), $\boldsymbol{\gamma}$ is the coefficient vector capturing the effects of lagged scores, $\mathbf{Z}_{qi}$ is a design matrix with one column for each unit in q ($q \in \{1,2,\ldots,Q\}$) and one row for each student record in the database. The entries in the matrix indicate the association between the test represented in the row and the unit (e.g., school, teacher) represented in the column. Sub-matrices are concatenated such that $\mathbf{Z} = \{\mathbf{Z}_1,\ldots,\mathbf{Z}_Q\}$. $\boldsymbol{\theta}_q$ is the vector of effects for the units within a level. All test scores are measured with error, and the magnitude of the error varies over the range of test scores.

$$y_{ti} = \mathbf{X}_i\boldsymbol{\beta} + \sum_{r=1}^{L} y_{t-r,i}\gamma_{t-r} + \sum_{q=1}^{Q} \mathbf{Z}_{qi}\boldsymbol{\theta}_q + e_i$$

Predictor Variables

The main independent variables in the models are teacher observation scores. Two different observation scores were calculated. Using observation data collected from the MTEM, we collapsed the elements to match the new competencies in the FTEM. This was done to serve as a proxy for the FTEM because this was the first year the framework was launched, so sample sizes were much larger in the classic model. For the MTEM observation score calculation, scores were averaged to each element, then averaged to the matching FTEM competency (60 elements collapsed into 23 competencies), then averaged to individual teachers, so each teacher had an average observation score. The teacher observation score from the FTEM dataset was derived directly from the scores collected from the standard FTEM form. For the FTEM teacher observation score calculation, scores were averaged to the competency and then to the teacher.

The FLDOE selected the MTEM as the state model for teacher evaluation, and teacher value-added measures are of public record. Many districts in Florida also use iObservation. iObservation is an instructional and leadership improvement system that collects and reports data from teacher evaluations, observations, and informal walkthroughs. Data were exported from iObservation for customers using standardized forms for the classic and reformed model. The MTEM dataset included eight large public-school districts with 7,490 teachers that had a matching value-added measure. The FTEM dataset consisted of nine smaller districts with 488 teachers that were matched to a value-added measure.

Observers or raters of the models can include principals, assistant principals, administrators,

coaches, or district personnel. Observers use a 5-point performance scale in either model to assess levels of implementation of the elements or competencies. Observers typically have received some type of training or technical assistance on the model they are scoring. However, because the amount of training and guidance an observer can receive can vary substantially by district or school observational system, some system-level characteristics are accounted for within the more robust hierarchical linear models. The observation system characteristics include: the number of elements/competencies scored across observations, the number of times a teacher was observed, and the number of raters that observed a teacher. Failing to specify these important criteria can impact the reliability of the observation score (Hill, Charalambouse, & Kraft, 2012).

The last variable included in the hierarchical linear model includes a measure for poverty in the geographic location of the school in which the teacher works. Per statutory requirement, the FLDOE does not control for poverty in the teacher value-added measure calculation. Thus, economic influences in different geographic areas could impact the study results (Ballou, Mokhur, & Cavalluzzo, 2012; Goldhaber & Hansen, 2010; Johnson, Lipscome, & Gill, 2014; McCaffrey, Koretz, Lockwood, & Hamilton, 2004; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Staiger & Kane, 2014; Stuit, Berends, Austin, & Gerdeman, 2014). While the inclusion of prior test scores (and other controls) accounts for some variation in the economic disparities across regions, to control for poverty we included a Distressed Communities Index created by The Economic Innovation Group (2017). The index includes seven factors of poverty: adults without a high school diploma, poverty rate, prime-age adults not in work, housing vacancy rate, median income ratio, change in employment, and change in establishment. The index was linked by zip code to the school in which the teacher works.

Findings

Teacher observation scores were matched with value-added measures by first and last name and district. Eight public school districts were included in the MTEM dataset including 7,490 teachers who had an observation and combined teacher value-added measure. For the FTEM dataset, nine public school districts were included with 488 teachers who had an observation score and combined teacher value-added measure. Teachers included in the study sample are only those who administer the ELA or math state assessment within districts that implement the corresponding models and who use iObservation in the state of Florida.

The magnitude of effects is in line with other observation rubrics which find small to moderate correlations between observation scores and teacher value-added measures (Basileo & Toth, 2019). They report correlation coefficients for three years of data, including up to 13,316 teachers. The coefficients were small, positive, and statistically significant ranging from .17 to .19 for the combined value-added measure, .15 to .27 for ELA, and .19 to .23 for math.

Correlation coefficients were investigated to assess the magnitude of the relationship between teacher observation scores and value-added measures (Basileo & Toth, 2019). Coefficients are classified in magnitude using Cohen's conventions (1988) to interpret effect sizes of .10, .50, and .80 standard deviations as small, moderate, and large. Table 1 reports the coefficients for both models. There were small, positive, and statistically significant correlation coefficients between the average teacher observation score and value-added measures regardless of the model.

The magnitude of effects is in line with other observation rubrics which find small to moderate correlations between observation scores and teacher value-added measures (Basileo & Toth, 2019). They report correlation coefficients for three years of data, including up to 13,316 teachers. The coefficients were small, positive, and statistically significant ranging from .17 to .19 for the combined value-add measure, .15 to .27 for ELA, and .19 to .23 for math. As in this study, the authors also found that coefficients tended to be larger in math than in ELA. While these correlation coefficients are still classified by Cohen's (1988) definition as small, they are substantively higher than noted in previous studies.

Table 1. Correlation Coefficients of Teacher Observation Scores and Value-Added Measures

		Combined Value-Added	ELA Value-Added	Math Value-Added
MTEM Elements Collapsed	Observation Score	.243**	.219**	.270**
	N	7,490	5,924	3,674
FTEM	Observation Score	.246**	.178**	.328**
	N	488	365	225

***All correlations are statistically significant at the p < .01 level*

Adjusted coefficients are presented in Table 2. The adjusted coefficients are calculated following a series of studies that correct for attenuation due to the unreliability in the predictor and outcome measures (Hunter & Schmidt, 1990; Hunter & Schmidt, 1994; Marzano, Walters, & McNulty, 2005; Spearman, 1904). To correct for attenuation, one divides the observed correlation by the square root of the product of reliability coefficients. In this case, the reliability of the observation score is .617 and it was calculated using the percent of agreement after independent coding from two studies using the MTEM (Marzano Research Laboratory, 2011; Marzano, Toth, & Schooling, 2012). There is yet to be a reliability study for the FTEM, so this should be an area for future research. The reliability of the Florida Standards Assessment (FSA) in the 2017–2018 school year was .93 (Florida Standards Assessments, 2018). The square root of the product of the two reliabilities equates to .758. Dividing the Pearson coefficients by square root of the product provides the correlation coefficient, which controls for error between the predictor and outcome variables. After correcting for attenuation, correlation coefficients increased but were small in magnitude. Further investigation is warranted to assess whether observation scores predict the teacher value-added measures controlling observation system-level characteristics and school-level poverty rates. Next, multicollinearity is investigated, then more robust multilevel models are presented.

Table 2. Correlation Coefficients Corrected for Attenuation

	Combined Value-Added	ELA Value-Added	Math Value-Added
MTEM Elements Collapsed	.321	.289	.356
FTEM	.325	.235	.433

Multicollinearity was investigated in both datasets. For the classic model dataset, all correlation coefficients between the outcome and predictors were less than .299. For the FTEM dataset, the number of observers and the number of observations conducted were highly correlated (.652). This is most likely due to the small number of districts implementing the new model as there has not been much time to train observers. As this was the first year of implementation, sample sizes are notably smaller than the number of teachers observed using the classic model. Because of the smaller sample sizes, only correlation analyses will be reported for the FTEM dataset; however, both datasets were investigated, and the findings remained unchanged regardless of which teacher observation score and/or dataset was employed.

For the classic model dataset, all correlation coefficients between the outcome and predictors were less than .299. For the FTEM dataset, the number of observers and the number of observations conducted were highly correlated (.652).

Table 3 shows the descriptive statistics utilized in the multilevel model. Multilevel modeling is necessary to control for the nesting of and nonrandom selection of teachers within schools. Failing to account for the non-independence of observations can result in standard errors that are biased downward, increasing the likelihood of making inaccurate conclusions (Raudenbush & Bryk, 2002). Multilevel models correct for the dependence of error terms by incorporating a unique random effect for each of the equations nested within upper-level hierarchies. The following variables, while not exhaustive, could impact both the predictor and outcome variables and were included in the multilevel model.

Table 3. Descriptive Statistics for the MTEM Study Sample

Variable Name	Mean	SD	Min	Max
Observation Score	3.34	0.45	0.00	4.00
Competencies	13.69	4.72	1.00	64.00
Observations	5.35	2.53	1.00	25.00
Raters	1.79	0.74	1.00	6.00
Distress Index	39.88	26.85	1.25	98.86
Combined VAM	0.073	0.47	-3.78	4.75

The MTEM study sample consisted of 7,490 teachers in Florida across 636 schools in eight districts during the 2017–2018 school year. The average observation score for teachers included in the sample was 3.34 and ranged from zero to 4.00. The mean combined value-added measure was .073 and ranged from -3.78 to 4.75. The three estimates used to control for variation within observational systems included: the number of competencies included in the observation score (mean = 13.69), the number of

observations completed throughout the school year (mean = 5.35), and the average number of raters that observed teachers (mean = 1.79). Last, the only level two variable included in the model was the Distressed Communities Index. The mean distress index score was the MEAN according to the table is 39.88 with scores ranging from 1.25 to 98.86. Higher values indicate more distress in that area.

Table 4. Observation Scores Predicting Teacher Value-Added

Variable Name	B	SE	P-value
Intercept	-.7147	.0544	.0000
Observation Score	**.2376**	**.0135**	**.0000**
Competencies	.0032	.0013	.0150
Observations	.0194	.0088	.0280
Raters	-.0122	.0027	.0000
Distress Index	-.0004	.0003	.1880

Table 4 shows the multilevel model which incorporates all predictor variables. The observation score is statistically significant and is the largest predictor in the model (coefficient = .2376). All predictors are statistically significant except for the Distress Index. The intraclass correlation coefficient (ICC) for the intercept only model (not shown) is .097. The ICC measures the degree of dependence among observations within schools. When the intercorrelation is close to zero there is little clustering at schools. However, to make sure a multilevel model is still needed, we followed Muthen and Satorra (1995) and calculated the design effect which was slightly over two (2.15), indicating the multilevel model was indeed necessary. While not shown here, the observation score remains the largest and statistically significant predictor of teacher value-added regardless of whether the ELA or math value-added measure is used as the outcome. These findings were also true when employing the same models (minus the number of observations due to multicollinearity) for the teacher observation scores collected from the FTEM.

The observation score is statistically significant and is the largest predictor in the model (coefficient = .2376).

Conclusion

The major finding from this study is that regardless of whether the classic or reformed model was employed, the FTEM shows higher correlation coefficients between teacher observation scores and value-added measures. Correcting for attenuation increased the magnitude of the correlations but coefficients remained small in magnitude. We also directly tested the association between teacher observation scores and value-added measures, using the classic model collapsed into the revised model competencies. The random intercept models showed that the observation score was the largest and statistically significant predictor of teacher value-added, regardless of the subject and even when using data collected from the standard FTEM form for the same year.

While the sample used in this study is large, there are several limitations to using such data. Although some of the variance in the practice of observing teachers was accounted for in this study, the characteristics used in the analysis were not exhaustive. Teacher characteristics have been shown to impact teacher value-added (Basileo & Toth, 2019) but these measures were not available for the current year of study. Furthermore, there are still many unknown systematic confounders that could impact the results and may be stronger predictors of teacher value-added measures, particularly around the level of training raters have received. Finally, other school-level confounders—such as school rates of teacher turnover, teacher absenteeism, and student mobility—were not accounted for and could impact the findings as they are related to student achievement (Bailey, Bocala, Shakman, & Zweig, 2016).

Despite these limitations, the study confirmed the predictability of the revised model by analyzing two datasets collected in a real-world setting. The FTEM was redesigned to save time for administrators and to promote more accurate scoring. While it is unknown whether the new design saves school leaders time, the findings from this study indicate that condensing the number of elements scored and having a more concise model provides greater predictability than the classic version. While there is the possibility that the correlation levels only increased due to changes in the state assessment, the evidence from the multilevel model tends to point toward the strength in the reformed model. These findings would be even more valuable to practitioners if there was evidence that the revised model also saves school leaders time. This study validated the use of the FTEM in Florida by upholding and surpassing the magnitude of correlations found in prior studies, and by demonstrating that observation ratings were the largest predictor in multilevel models.

Policy Implications

Closing student achievement gaps remains a national priority. The ESSA legislation calls for fairer evaluation and support systems that are based on evidence of student achievement. The law calls for developing and disseminating high-quality evaluation tools, such as classroom observation rubrics, and to have evaluation results inform decision-making about professional development and improvement strategies. Change is imperative for observational frameworks to be less time consuming and more accurate.

In response to the ESSA legislation, and in addition to the demand from school leaders to have a more concise evaluation model, the MTEM was redesigned to be more succinct, measurable, and to have a greater focus on improving teacher performance. The reformed model was tested to assess whether the more concise framework could uphold the level of evidence needed for teacher evaluation models to be predictive of student achievement. The findings here validate the use of the FTEM, particularly in Florida, by upholding the magnitude of correlations found in other instructional frameworks and by surpassing those in other studies on the classic model. Additionally, the study found that observation ratings were the largest predictor of teacher value-added accounting for student, observational characteristics, and poverty. While it is unknown whether the new design saves school leaders time, the findings from this study indicate that condensing the number of possible scores a teacher can receive and having a more concise model does provide greater predictability than the classic version. If future research could provide evidence the reformed model saves school leaders time, it would be invaluable.

Teacher evaluation makes up a small part of a principal's overall responsibilities, but it is an essential aspect of school leadership. Evaluation models must have accuracy in scoring, a predictable effect on student success, and a time-efficient design. This study demonstrates the FTEM fulfills the need for accuracy and predictability, and its streamlined design has the promise of reducing overall time investment. In turn, more simplified frameworks can better support administrators conducting teacher evaluations. The extent to which other frameworks may need to be reformed to meet the demands of the ESSA is unknown. Redesigning evaluation frameworks needs to be done wisely—without sacrificing the level of evidence needed to accurately score, provide feedback, and predict student achievement. The more evaluation frameworks can meet the demands of practitioners, while still upholding strong levels of evidence, the better they will meet the demands of the ESSA. While more research is necessary, the evidence provided in this study yields promise to meet the demands put forth in the ESSA legislation.

References

Alexander, S. (2016). *The relationship between teacher evaluation model, value-added model, and school grades.* (Doctoral dissertation). ProQuest Dissertations and Theses Global database. (UMI No. 10141740).

Alger, V. E. (2012). Teacher selection and evaluation in Nebraska. https://www.platteinstitute.org/Library/docLib/20120109_Teacher_Selection_and_Evaluation_in_Nebraska.pdf.

Auguste, B., Kihn, P., & Miller, M. (2010). Closing the talent gap: Attracting and retaining top third graduates to careers in teaching. http://mckinseyonsociety.com/-closing-the-talent-gap/.

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher demographics and evaluation: A descriptive study in a large urban district* (REL 2017–189). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. http://ies.ed.gov/ncee/edlabs.

Ballou, D., Mokher, C. G., & Cavalluzzo, L. (2012). Using value-added assessment for personnel decisions: How omitted variables and model specification influence teachers' outcomes. Paper presented at the annual meeting of the Association for Education Finance and Policy. http://www.aefpweb.org/sites/-default/files/webform/AEFPUsing%20VAM%20for%20personnel%20decisions_02-29-12.pdf.

Basileo, L. D. (2016). Truth or Myth About the Marzano Teacher Evaluation Model. Learning Sciences Marzano Center. https://www.learningsciences.com/wp/wp-content/uploads/2018/05/Truth-or-Myth-About-the-Marzano-Teacher-Evaluation-System.pdf.

Basileo, L. D., & Toth, M. (2019). A State Level Analysis of the Marzano Teacher Evaluation Model: Predicting Teacher Value-Added Measures with Observation Scores. *Practical Assessment, Research and Evaluation.* https://openpublishing.library.umass.edu/pare/article/id/1586/.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching. Measures of effective teaching.* http://k12-education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/.

Carbaugh, B., Marzano, R. J., & Toth, M. D. (2017). The Marzano Focused Teacher Evaluation Model: A Focused, Scientific-Behavioral Evaluation Model for Standards-Based Classrooms. https://www.learningsciences.com/wp/wp-content/uploads/2017/06/Focus-Eval-Model-Overview-2017.pdf.

Childress, M. (2014). Tools to strengthen instruction through multiple measures of evaluation. *Principal.* May/June. https://www.naesp.org/sites/default/files/Childress_MJ14.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Elementary and Secondary Education Act (ESEA) of 1965 [As Amended Through P.L. 115–224, Enacted July 31, 2018] https://legcounsel.house.gov/Comps/-Elementary%20And%20Secondary%20Education%20Act%20Of%201965.pdf.

Florida Value-Added Model Technical Report (2013). American Institutes for Research. http://myflteacher.com/wp-content/uploads/2014/05/Value-AddedModelTechnicalReport1213.pdf.

Florida Standards Assessments. (2018). *Florida Standards Assessments 2017-18: Volume 4 evidence of reliability and validity.* The Florida Department of Education. http://www.fldoe.org/core/fileparse.php/5663/urlt/V4-FSA-1718-TechRpt.pdf.

Goldhaber, D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance* (CEDR Working Paper 2010-3). Center for Education Data and Research, University of Washington. https://www.cedr.us/papers/working/CEDR%20WP%202010-3_Bad%20Class%20Stability%20(8-23-10).pdf.

Grissom, J. A., Loeb, S., & Master, B. (2013). Effective Instructional Time Use for School Leaders: Longitudinal Evidence from Observations of Principals. *Educational Researcher, 42*(8), 433-444.

Haystead, M., & Marzano, R. (2009). *Meta-analytic synthesis of studies conducted at Marzano Research Laboratory on instructional strategies.* https://www.marzanoresearch.com/meta-analytic-synthesis-of-studies.

Hill, H., Charalambos, Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Sage.

Hunter, J., & Schmidt, F. (1994). Correcting for sources of artificial variable across studies. In H. Cooper & L. Hedge (Eds.). *The handbook of research synthesis* (pp. 323-336). Russell Sage Foundation.

Johnson, M., Lipscomb, S., & Gill, B. (2014). Sensitivity of teacher value-added estimates to student and peer control variables. *Journal of Research on Educational Effectiveness, 8*(1), 60–83.

Marzano, R. J. (2006). *Classroom assessment & grading that work.* Association for Supervision and Curriculum Development.

Marzano, R. J. (2007). *The Art and Science of Teaching: A Comprehensive Framework for Effective Instruction.* Association for Supervision and Curriculum Development.

Marzano, R. J., Marzano, J. S., & Pickering, D. J. (2003). *Classroom management that works: Research-based strategies for every teacher.* Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D.J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement.* Association for Supervision and Curriculum Development.

Marzano, R. J., Toth, M., & Schooling, P. (2012). *Examining the role of teacher evaluation in student achievement*. http://ok.gov/sde/sites/ok.gov.sde/files/TLE-MarzanoWhitePaper.pdf.

Marzano Research Laboratory. (2011). *What works in Oklahoma schools: A comprehensive needs assessment of Oklahoma schools: Phase 2 report.* https://sde.ok.gov/sites/ok.gov.sde/-files/SI-PhaseIIStateReport.pdf.

Marzano, R., Waters, T., & McNulty, B. (2005). *School leadership that works: From research to results.* Association for Supervision and Curriculum Development.

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267-316.

Newton, A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. http://epaa.asu.edu/ojs/article/view/810.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115 Stat. 1425. (2002). http://www.ed.gov/legislation/ESEA02/.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101.

Staiger, D., & Kane, T. (2014). Making decisions with imprecise performance measures: The relationship between annual student achievement gains and a teachers' career value added. *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (pp. 144–169). Jossey-Bass.

Stuit, D., Berends, M., Austin, M., & Gerdeman, D. (2014). *Comparing estimates of teacher value-added based on criterion- and norm-referenced tests.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. http://ies.ed.gov/-ncee/edlabs/projects/project.asp?ProjectID=392.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: *Applications and data analysis methods.* Sage.

Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on school outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*(5), 635-674.

Renter, D. (2012). *After the stimulus money ends: The status of state K-12 education funding and reforms.* http://www.cep-dc.org/displayDocument.cfm?DocumentID=395.

The Economic Innovation Group. (2017). *The 2017 Distressed Communities Index.* https://eig.org/dci.