Examining the Association Between Teacher Annual Evaluations and Student Academic Performance: Recent Outcomes for North Carolina

Theodore Kaniuka, Fayetteville State University
Brad Mills, Fayetteville State University
Ashley Johnson, Mount Olive University
Emily Haire, Fayetteville State University

Corresponding author: Theodore Kaniuak
Email: tkaniuka@uncfsu.edu

Measuring teacher effectiveness has been debated and studied for numerous years. While some progress has been made, consensus has yet to be reached regarding what it means to be an effective teacher and how to measure effectiveness. This study uses administrative data from North Carolina to assess the relationship between school principal evaluations of teachers and student achievement based on a value-added measure. Multi-level linear regression results suggest that 1) teachers account for the majority of the variation in value-added scores, 2) principal evaluations of teachers have some predictive value relative to the teacher effectiveness measure, and 3) teacher evaluation scores fail to signal teacher effectiveness scores. Outcomes are discussed in terms of educational leadership and policy.

Keywords: teacher evaluation, educational measurement, value-added

Examining the Association Between Teacher Annual Evaluations and Student Academic Performance: Recent Outcomes for North Carolina

Introduction

In the State of North Carolina, teacher evaluations aim to document professional performance and an array of teaching behaviors and serve as a foundation for the personal growth of teachers (North Carolina Educator Effectiveness System [NCEES], 2018). Teacher evaluation in North Carolina has been required by statute to support teacher development and as a method to support administrative actions when necessary (GS 115c-333-333.2, 2022). Recent revisions to the North Carolina process included the introduction of the North Carolina Educator Effectiveness System (NCEES, 2018). The NCEES was developed in response to the requirements in the Race to the Top (RTTP) legislation (American Recovery and Reinvestment Act, 2009), which requires that education agencies develop systems to measure and document teacher effectiveness. As currently implemented, the reliability and validity of the NCEES process scores are reliant upon trained administrators who consistently apply evaluation criteria and procedures. These administrators evaluate teachers using an instrument that captures teaching practices to provide valuable performance information to both teachers and administrators. One of the key elements to teacher evaluation and, subsequently, communication of effectiveness is the use of data from the Education Value-Added Assessment System [EVAAS] (North Carolina Department of Public Instruction [NCDPI], 2024). This is a measure of students' academic growth for accountability purposes and is attributed to the student's teacher of record. The system uses individual and statewide yearly student achievement data from the state's annual testing program to calculate the student's growth in academic performance. It is often called a value-added measure as it links individual student growth performance to a teacher. It is defined as the value in the growth of educational attainment a student experiences due to having a particular teacher. It allows for documenting teacher instructional effectiveness and school performance and is used to assign annual school performance grades and monetary awards to educators (NCDPI, 2024).

This study builds on prior work linking principal evaluations to value-added student growth measures (Brophy, 2010; Darling-Hammond, 2010; Harris & Sass, 2007; Jacob & Lefgren, 2008). These studies and others (Stronge et al., 2011) found that principal evaluations can predict student outcomes, but such measures lacked consistency and reliability. Adding to earlier work, a recent study involving District of Columbia teachers by Dee et al. (2021) found that principal evaluations coupled with meaningful accountability can improve the quality of teaching. This was consistent with the findings of the Measures of Effective Teaching [MET] (Center for Education and Policy Research, 2013) project. The MET found that principal evaluations of teachers were significant predictors of student outcomes when the process was well designed, used valid instruments, and had enough variation in teacher ratings to reflect the diversity in teacher skill and performance. Considering the above, the overarching aim of this study was to examine how well the results from the teacher evaluation process predict student outcomes and if the evaluation results could be applied to support educational decision-making. The following research questions were used to structure the inquiry:

1) How well do teacher annual evaluation scores predict value-added[1] measures?
2) Do teacher evaluation scores signal teacher value-added scores?

[1] In North Carolina value added is measured by Education Value-Added Assessment System (North Carolina Department of Public Instruction [NCDPI], 2024)

Teacher Evaluation- Theory

Teacher evaluation can be supported from two contrasting theoretical perspectives. The first emphasizes intrinsic motivation: teachers are motivated to receive and participate in evaluations to support professional development and improve overall teaching performance. Intrinsic motivation occurs when the teacher sees value in the activity and the rewards are self-fulfilling (i.e., Ryan & Deci, 2000). A supporting theory is self-efficacy (Bandura, 1977), where teachers receive feedback and experiences that help them improve their self-confidence in the task at hand. This feedback can take the form of direct experiences and other sources that provide teachers with a sense of ability or belief that they can be successful or influence the actions of others (Tschannen-Moran et al., 1998). These theories trace their roots to Vroom (1964) and apply the concept of expectancy, where teachers believe that their efforts have an intended outcome, such as feeling that they have accomplished a goal or added value to the student's lives.

In contrast, extrinsic motivation is receiving some external recognition of a job well done. Intrinsic motivation is engaging in activities solely for the rewards or degree of satisfaction those activities provide to the individual. Extrinsic motivation is engaging in activities with the expectation that some form of reward outside the individual will be provided (see Morris et al., 2022). In the case of contemporary teacher evaluation systems, extrinsic motivation may take the form of annual performance bonuses or merit pay. The North Carolina School Accountability System (NCDPI, 2024) adheres to this notion, as teachers who show acceptable student growth receive a bonus. This idea of rewarding teachers is consistent with federal policies that reward and punish teachers for performance (i.e., Every Student Succeeds Act, 2024). This current study takes these two competing views of teacher evaluation by examining the link between formal teacher evaluation and student performance. The NC system is designed to support professional growth and development based on a comprehensive annual evaluation process (NCEES, 2024, North Carolina Teacher Evaluation Process, 2015) and an accountability system that rewards and punishes based on teacher performance as measured by student outcomes.

Teacher Effectiveness

The concept of teacher effectiveness has been the subject of academic research for nearly 70 years, and this research has established a clear link between teacher behaviors and student outcomes (see Goe et al., 2008; Skourdoumbis, 2014; Slater et al., 2012). However, there are unresolved issues regarding an effective teacher and how we can measure effectiveness. Measuring teacher effectiveness is not only complicated by perspectives on the defining characteristics of an effective teacher (Berliner, 1976; Campbell et al., 2003; Cheng & Tsui, 1999; Cruickshank & Haefele, 1990; Good, 1996) but confounded by methods used to measure this construct (Holmes & Schumacker, 2020; Garnett, 2013; Garrett & Steinberg, 2015; Muijs, 2006). According to Goe et al. (2008), effective teaching is more than simply using one or two measures to quantify behaviors. Instead, it may be context-specific and require a more holistic data set to portray teacher effectiveness accurately. For example, Walker (2020) listened to students speak about teachers, revealing twelve characteristics of effective teachers. Students stated that effective teachers are those who, among other things, demonstrate a) preparedness, b) having high expectations, c) relating on a personal level, d) showing compassion, and e) providing respect. Students see effective teachers as more than just being prepared or knowledgeable; they value interpersonal skills and a sense of humanity as essential qualities of an effective teacher.

To support the RTTP legislation and develop stakeholder-driven effectiveness systems, the MET (Center for Education and Policy Research, 2013; Kane & Staiger, 2012) project spent three years investigating various teacher evaluation models and supporting research on measuring teaching. The project not only yielded a set of recommendations upon which states and school districts could design a teacher evaluation process but also found that the current systems in use needed to be revised. The systems of teacher evaluation studied as part of the project were found to be defective to the extent that the MET project discovered that nationally, approximately 98% of teachers are rated as satisfactory. However, schools routinely fail to have student performance outcomes that align with these ratings. This degree of consistency and associated absence of substantial variation in teacher performance ratings belies the variation of student performance both within and between schools. The MET's culminating report presented recommendations for a more robust evaluation process that includes classroom observations, student surveys, and student achievement gains as a holistic system to capture the various aspects of teaching better.

In North Carolina, a report (Race, 2010) to the State Board of Education mirrored the recommendations offered by the MET project and influenced the NCEES design. The main finding and recommendation of the report emphasized North Carolina's intent to combine a value-added measure along with five additional measures of the work of teachers to create a more holistic evaluation process. Seemingly consistent with the perspective shared by Dee et al. (2021), the North Carolina approach could be linked to school improvement work or used as a reliable and valid predictor of school performance.

North Carolina Educator Effectiveness System

With the NCEES, North Carolina has been using teacher evaluation data to quantify teacher behaviors, develop teacher annual growth plans, and document the effects of teachers' effectiveness on student performance. This has been accomplished by including principal observation data and a value-added measure (NCEES, 2018). According to the most current version of the North Carolina Teacher Evaluation Process (North Carolina Department of Public Instruction [NCDPI], 2018, p5), the process has eight distinct aims. Among them are: a) serve as a measurement of the performance of individual teachers; b) serve as a basis for instructional improvement; and c) focus on the goals and objectives of schools and districts as they support, monitor, and evaluate their teachers.

The teacher evaluation system was originally a rubric-based approach utilizing six standards: T1) Teachers demonstrate leadership; T2) Teachers develop a respectful environment for a diverse population of students; T3) Teachers know the content they teach; T4) Teachers facilitate learning for their students; T5) Teachers reflect on their practice; and T6) a value-added outcome. The source of the value-added component was the Education Value-Added Assessment System (EVAAS), used by the North Carolina Department of Public Instruction (NCPDI, 2024) to measure student growth achieved during the academic year. EVAAS was developed to assess teacher performance and the use of instructional programs and approaches by analyzing student test performance over time to quantify the impact the teacher and these programs had on student outcomes. Using inferential statistical methods, EVAAS attempts to account for the influence of alternative contributing factors and allows evaluators to isolate the teacher's contribution to individual student academic gains (Vosters et al., 2018). This process has been criticized as being too complex, not transparent enough, lacking validity, and that more straightforward methods yield results equivalent to or outperform the methods in EVAAS (EdNC, 2017; Vosters et al., 2018).

The five standards, T1 through T5, are measured using a 5-point scale defined as 1) Developing, 2) Proficient, 3) Accomplished, 4) Distinguished, and 5) Not Demonstrated. The principal assigns scores due to direct classroom observations and a review of professional artifacts (i.e., instructional lesson plans and professional development plans). The Not Demonstrated rating is reserved for situations where the principal did not observe the desired behaviors. To support principals, an observation rubric is used to determine each teacher's performance rating, as each level of performance has sample criteria with which the principal can anchor their ratings and assign a score. After two years of including the value-added standard as part of the evaluation process, it was removed in response to concerns about using student performance as an evaluation component. Although it was removed as an official aspect of the annual evaluation process, each teacher continues to receive an EVAAS score used to award bonuses, and each school receives an aggregate EVAAS score.

Both teachers and principals have documented responsibilities. Of those, the two most closely aligned to this study are: (a) teachers gather data, artifacts, and evidence to support performance about standards and progress in attaining goals, and (b) principals ensure that the contents of the Teacher Summary Rating Form contain accurate information and accurately reflect the teacher's performance.

Teachers are responsible for demonstrating artifacts and evidence of adherence to rubric expectations and requirements. This includes the professional development plan collaboratively developed with the principal to reflect the previous evaluation outcomes and the teacher's professional growth goals. There are no delineated requirements that a professional development plan address student learning outcomes unless the teacher is on a directed growth plan to address ratings of *developing* or *not demonstrated* or for the value-added score of the teacher to be considered *not met*. Therefore, it is entirely incumbent upon the principal and teacher to collaboratively develop a growth plan for an individual teacher's growth and career path. Initially, all principals and assistant principals receive guided professional learning on the NCEES process. Once trained, principals and assistant principals meet to review the evaluation process, where updates and revisions are shared; however, there is no state-wide systematic and systemic learning to support the maintenance of the validity and reliability of the observation process once the initial training is completed. Self-directed teacher training is provided via a web-based portal sponsored by the North Carolina Department of Public Instruction.

Data and Methods

The data analyzed in this study were from North Carolina administrative teacher-level files for the 2016 and 2017 school years. There were approximately 65,000 teachers in the data set, and 51,194 teachers had valid EVAAS scores. The remaining teachers in the data set did not have an EVAAS score or taught a subject not part of the state testing program. For each teacher in the study, these data contained the NCEES scores for the five standards, the EVAAS index score, and the covariates of years of experience, race, and sex. It was hypothesized that the school the teacher was assigned to could influence the relationship between the independent and dependent variables. Therefore, a multi-level linear regression approach was used to cluster teachers into their assigned schools to account for the between-school variance. The original 5-point scale used in the NCEES had a measure indicating *Not Demonstrated*. This score point is used to communicate that the evaluated teacher did not have or supply evidence for the behavior from formal observations or other sources such as lesson plans, professional learning, or other data types. This score point was seen as problematic as not supplying any evidence is not necessarily an indication of poor performance; instead, it was coded that the teacher chose not to furnish the required data or that the principal did not observe the target behavior. The teacher may have been retiring, opted to move to another position, or was being removed. Given the ambiguity associated with this score point, it was decided that including it in the analysis may bias the results. Examination of the data showed that approximately 8.5% of the sample received this score, resulting in including approximately 91% of the original sample in the analysis.

The study's design utilized archival data generated for other administrative purposes. The use of observational data has limitations, and it has been argued that it is inappropriate in studies, namely when the researchers attempt to compare groups (Rosenbaum et al., 2010). This study does not attempt to conduct a causal-comparative analysis but instead utilizes a cross-sectional correlational approach, recognizing that this design is limited in scope (Lau, 2017; Leigh, 2010) and that conclusions can be developed. To that end, this study sought to understand if there was a relationship between the scores principals assigned as part of an evaluation process and teacher effectiveness scores and if that relationship can be used to link annual evaluation scores to the idea of teacher effectiveness. Additionally, if a relationship was found, could that relationship be applied as a screening device for administrative purposes? The utilization of a linear mixed model regression was selected as it was hypothesized that there may be a variance component attributable to clustering teachers within schools (i.e., Muñoz et al., 2011; Subedi, 2015). Using a random intercept linear regression provided the opportunity to determine the variation across schools that was not captured by the fixed effects portion of the model. Including a limited number of covariates in the model helped to account for some of the fixed effects associated with teachers, such as race, experience, and sex. The general form of the model is shown below.

$$yij=\beta 0+\beta 1Xij+uj+\epsilon ij \qquad (1)$$

Where y_{ij} is the EVAAS score for each teacher *i* in a particular school *j*, β_0: the fixed intercept, $\beta 1x_{ij}$: the fixed effects vector of the predictor variable (NCEES scores), $\beta 2x_{ij}$: is a vector of covariates; u_{0j}: the random intercept for school, and ϵ_{ij}: The residual error for individual *i* in school *j*. The NCEES scores were modeled as indicator variables, with the score point 3 (a *Proficient* rating) being the reference category. This allowed the researchers to compare the estimates above and below *Proficient*, the North Carolina NCEES evaluation target score. Before data analysis, EVAAS scores were standardized by test type to adjust for the differences in test metrics. In grades 3-8, students take a reading, math, and science test developed by the state. In high school, students take English, Biology, and math exams. Other courses have standardized final examinations across the state but differ from the other accountability exams as they are not used to calculate the percentage of students considered to meet state proficiency standards. Instead, they have an accountability measure as required by state policy. In K-2, state exams are not permitted; therefore, reading and math data are gathered when teachers conduct individual assessments of children.

The unbalanced clustering of teachers in schools was one of the utmost concerns as increasing cluster size has benefits such as improving the power to estimate random effects (Austin & Leckie, 2018). Although larger cluster sizes may be desirable, small clusters were found not substantially bias the estimation results (Clarke & Wheaton, 2007; Maas & Hox, 2005). A simulation study found that small sample sizes per cluster unit are not problematic (Bell et al., 2008) if the number of clusters is approximately 500 or larger. Thus, having cluster sample sizes equaling one is not associated with estimation and bias issues; therefore, linear mixed modeling was used. This study had 2481 schools or clusters that far exceeded those recommended by Bell et al.

Before running the regression, a Kruskal-Wallis test was run to estimate the variance explained in EVAAS by the scores for each NCEES standard area. This was considered most appropriate as the NCEES scores are on an ordinal scale of 2-5 based on a rubric. The results indicated that all the chi-square statistics were highly significant for each standard analyzed, implying that the NCEES scores were associated with EVAAS. A multi-level regression was run, and it was anticipated that this approach would reveal if the different scores teachers received accounted for the variance in the EVAAS score while controlling for the influence of the covariates and the clustering of teachers in schools. The fundamental question was, when a teacher received a higher/lower evaluation score, did this score predict higher/lower EVAAS scores? As mentioned above, the *Proficient* rating was used as the reference score to which the other ratings were compared. This referencing provided an anchor that would reveal that if a teacher was rated below *Proficient*, it is reasonable to assume that such ratings predict lower EVAAS scores. Similarly, if a teacher received an *Accomplished* or *Distinguished* score and the estimated coefficient was positive, it would be reasonable to assume that such ratings would predict higher EVAAS scores.

Results

The results are first reported for the summary statistics for the independent and dependent variables, followed by regression estimates.

Summary Statistics

As seen in Table 1, the mean scores for each of the five NCEES standards are well above 4, as reported on a Likert scale ranging from 2 to 5. When looking at the distribution of scores from a different lens, it was found that over 97 percent of the teachers received a satisfactory score of *Proficient* or greater across the five standards. The EVAAS scores are split into three aggregate levels: did not meet, met, and exceed. There is a range associated with each score, and the met category ranges from ±1.99 in value. The score of -0.11 shows that, on average, students met the performance target. Within the sample of teachers in this study, 62.41 percent achieved Met, and 20.58 percent achieved Exceeds.

Table 1

Summary statistics for focal predictor and dependent variables (n = 51194 teachers)

Variable	Mean	S.D.	Scale
NCEES Evaluation Standard			
Teacher Leadership	4.435	0.984	Likert
Respectful Environment	4.462	0.533	
Content Knowledge	4.374	0.548	
Facilitate Learning	4.407	0.538	
Reflective Practice	4.373	0.549	
	Dependent		
EVAAS			
Standardized	-0.11	0.984	Continuous

Table 2 shows that the sample was highly female and that white teachers comprised approximately 80 percent. Regarding experience, the mean experience was less than ten years, with an M = 8.832 and an SD = 8.09. This indicates that in the years 2016-17, approximately 68 percent of teachers had between slightly more than a few months and 16 years of experience.

Table 2

Summary statistics for covariates (n= 51194)

Variable	Mean	S.D.	Scale
Teacher Experience	8.832	8.09	Continuous
Sex* (Females)	81.93	0.385	Categorical
Ethnicity*			
Black	16.22	0.368	Categorical
Hispanic	1.99	0.139	
Indigenous	1.08	0.1	
Other	0.1	0.043	
White	79.69	0.402	

*Note: State assigned codes, reported in percents, rounding errors not summing to 100 percent

Regression Results

The regression analysis results are reported in Table 3; the model was found to be significant as Wald $\chi^2(22) = 1838.31$, $p<0.001$ with n = 51194 clustered into 2481 schools. Using a multi-level model appears justified as the variance estimates for the level two clustering variable are both significant, and the interclass correlation coefficient was 0.125 and significant (Raudenbush & Byrk, 2002).

Table 3

Multi-level Regression on EVAAS by Teacher NCEES Rubric Areas

Rubric Area	Rating	Coefficient	Robust S.E.	z	p
Teacher Leadership					
	Developing	0.345	0.260	1.33	0.185
	Accomplished	0.037	0.042	0.87	0.386
	Distinguished	0.131	0.044	2.98	0.003
Respectful Environment					
	Developing	-0.190	0.291	-0.65	0.513
	Accomplished	0.215	0.045	4.76	<0.001
	Distinguished	0.263	0.047	5.66	<0.001
Content Knowledge					
	Developing	0.368	0.323	1.14	0.254
	Accomplished	-0.059	0.036	-1.62	0.105
	Distinguished	0.067	0.039	1.73	0.084
Facilitate Learning					
	Developing	0.163	0.188	0.87	0.387
	Accomplished	0.255	0.041	6.16	<0.001
	Distinguished	0.405	0.044	9.21	<0.001
Reflective Practice					
	Developing	0.006	0.237	0.03	0.979
	Accomplished	0.051	0.041	1.25	0.211
	Distinguished	0.102	0.042	2.41	0.016
	Constant	-0.874	0.067	-13.08	<0.001
Covariates					
Teacher Experience		-0.005	0.001	-7.31	<0.001
Sex		0.110	0.014	7.70	<0.001
Race					
	Black	-0.064	0.054	-1.18	0.236
	Hispanic	0.008	0.060	0.13	0.893
	Indigenous	0.017	0.071	0.23	0.815
	Other	-0.150	0.116	-1.29	0.196
	White	0.025	0.053	0.47	0.641

Random Effects

Variable	Source	Estimate	Robust S.E.	95% Confidence Interval	
				LCI	UCI
School	Variance (constant)	0.115	0.006	0.104	0.128
	Variance (Residual)	0.804	0.011	0.782	0.826

Residual Interclass Correlation

	ICC	Std Error	95% Confidence Interval	
Source				
School	0.125	0.006	0.115	0.137

The results displayed in Table 3 show the predicted associations between NCEES ratings and EVAAS performance. Of the 15 predictions, 6, or slightly more than one-third, were found to have estimates statistically different from zero. Not one standard had all three estimates being significant. When considering all estimates, only a few had negative or positive values consistent with the hypothesis that rating below or above *Proficient* should have negative and positive estimates, respectively. Of note, all the significant estimates had values consistent with the hypothesis that higher ratings should be associated with higher EVAAS estimates.

Two NCEES standards appear closely related to student learning outcomes as they examine 1) a teacher's knowledge about their curriculum and 2) whether they can create learning environments conducive to learning. Standard III – *Content Knowledge* assesses if the teacher understands the state curriculum content (North Carolina Standard Course of Study), can make connections across subjects, and makes instruction relevant and appropriate. As seen by the results presented in the table, not one of the estimated coefficients was found to be significantly different from zero, which was unexpected. It is the only standard that did not have at least one estimate being statistically different from zero, which means that compared to the *Proficient* rating, a rating less than or conversely greater than was not estimated to have a significant differential association with EVAAS scores. Standard IV- *Facilitate Learning* evaluates teachers if they can demonstrate that they know how learning takes place (how students learn) and can align their teaching to student needs (differentiated learning).

Additionally, this standard examines instructional planning, alignment of teaching and assessment to student needs, and technology integration. The fixed effect estimates show that the higher the ratings of Accomplished and Distinguished are significantly different from zero and positive. These estimates indicate that teachers rated as accomplished were predicted to have

EVAAS estimates 0.255 standard deviations higher than those rated as P*roficient. S*imilarly, *Distinguished* were predicted to have EVAAS scores 0.405 standard deviation units higher.

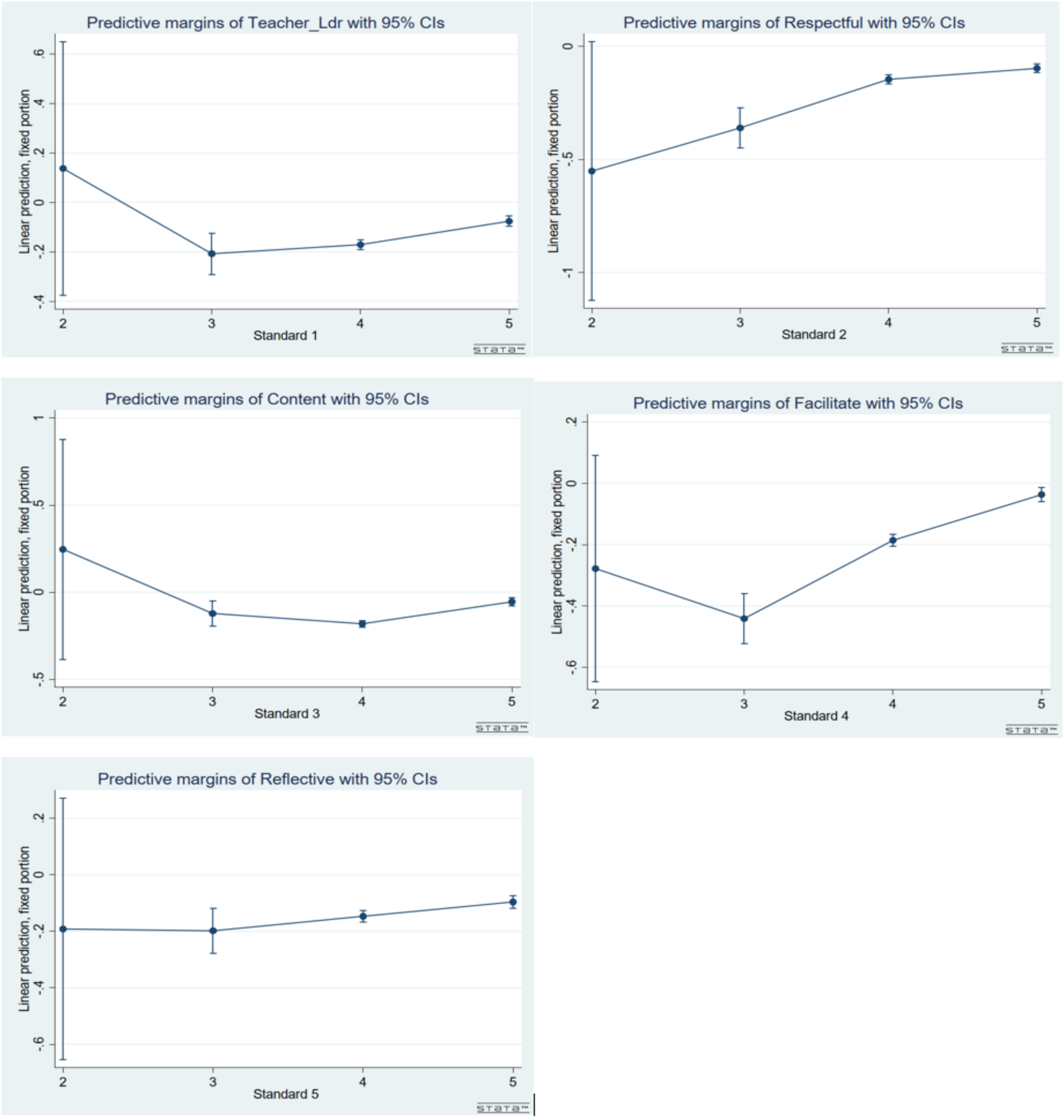
Findings for Standard II - *Respectful Environment,* which addresses such areas as teachers developing and providing a nurturing and respective environment, embracing diversity, and adapting their instruction to benefit students with special needs, revealed weak associations. Two of three estimates were significant, and these estimates were consistent with the hypothesis as the estimate for the accomplished and *Distinguished* levels indicated that the EVAAS scores are to be 0.215 and 0.263 standard deviation units higher than those rated at *Proficient,* which is greater, respectively.

Regarding the random portion of the model, the variance $\sigma_u^2 = 0.115$ is significant, as indicated by the values for the confidence intervals (LCI = 0.104, UCI = 0.128). This value for the variance reveals that the school a teacher is assigned to matters as it influences EVAAS scores. The second variance component $\sigma_e^2 = 0.804$, (LCI = 0.782, UCI = 0.826) shows that a much greater amount of the variance in baseline EVAAS scores is attributable to the teacher level portion of the model and that this variation is significant.

To understand each standard's characteristics, standard predictive margins were calculated. In this case, the predictive margins communicate the estimated EVAAS scores for each standard, holding constant the influence of the other standards. Calculating the margins for each standard shows only the predicted EVAAS scores. Each standard has four levels, and an EVAAS score was predicted for each, communicating the estimated score as a function of the levels: *Developing, Proficient, Accomplished, and Distinguished*. The results are presented below in illustration 1.

Illustration 1

Predictive Margins for NCEES Standards of EVAAS Scores



All the level 2 ratings were insignificant and provided little information on the NCESS standards' predictive nature. According to the margins analysis the remaining levels all provided significant estimates of EVAAS scores. In four of the five cases shown above, *Proficient* scores predicted smaller EVAAS scores as compared to *Developing*. Furthermore, for *Content Knowledge* this downward trend was seen for *Accomplished* as well. The standards of *Respectful, Facilitating*, and *Reflective* all saw the *Distinguishe*d level predicting the highest EVAAS scores. The inconsistent results from the margins analyses are indicating that the usefulness of the NCEES scores as they relate to student outcomes is worth considering.

Given the effort principals and teachers expend in the evaluation process and the inconsistent results from the regression analysis, a second research question emerged asking if evaluation results provide a valid signaling of teacher effectiveness? The ability of the NCEES rubric scores to screen effective teachers (as measured by the EVAAS scores) was assessed by estimating the probabilities that a teacher performs in the lowest and highest quintiles (the bottom and top 20% of EVAAS scores). To accomplish this, rubric scores were recoded into a binary variable indicating if a teacher was not *Proficient* or *Proficient*. A proportion test was used incorporating clustering and intraclass correlation (see Goldhaber et al., 2017). The rationale is that if the NCEES scores for each standard were ineffective screeners, we expect that 20% of the teachers who did not receive a *Proficient* rating would be distributed equally in the categories.

The results of this analysis are reported in Table 4. It was found that for those teachers not rated as *Proficient* approximately 10 percent were found to be in the lowest quintile for each of the five standards. Alternatively in some cases those rated as *Proficient*, had EVAAS scores in the lowest quintile.

Table 4

Quintile Proportions of Teacher Effectiveness for Proficient and Not Proficient Ratings by Evaluation Standards

Standard	Proficient	Quintiles			
		Lowest		Highest	
		Proportion	S.E.	Proportion	S.E.
Teacher Leadership	No	0.103***	0.005	0.309***	0.009
	Yes	0.209*	0.004	0.189**	0.004
Respectful Environment	No	0.115***	0.007	0.282***	0.009
	Yes	0.207	0.004	0.193	0.004
Content Knowledge	No	0.106***	0.007	0.323***	0.01
	Yes	0.207	0.004	0.191**	0.004
Facilitate Learning	No	0.1***	0.007	0.329***	0.011
	Yes	0.205	0.004	0.193*	0.004
Reflective Practice	No	0.106***	0.006	0.299***	0.01
	Yes	0.207*	0.004	0.192*	0.004

Note: Each cell shows the proportion of teachers not receiving/receiving proficient rating in each bottom and top quintiles for the EVAAS. Test of significance is against the null hypothesis that the proportion = 0.2. † p < 0.1, * p < 0.5. ** p < 0.01, *** p < 0.001

Additionally, when considering all the NCEES standards, teachers rated as below *Proficient* were over-represented in the highest quintile. For example, regarding the *Content Knowledge* standard, about 32.3 percent of teachers rated as not *Proficient* scored in the highest quintile which is greater than the 20 percent expected. Interestingly, teachers who were rated as being *Proficient*, had smaller proportions in the highest quintile when compared to those rated as not being *Proficient*. Stating this differently, if employment decisions were based on these evaluation ratings, approximately 30 percent of the so-called not *Proficient* teachers had some of the highest EVAAS scores, and this was consistently found for each of the four additional standards. This over representation of non-*Proficient* teachers having the highest EVAAS scores raises the question of the usefulness and accuracy of NCEES scores in terms of student outcomes. NCEES was a screener, however it screened the teachers in ways that appears to be inconsistent with logic.

Discussion

Much has been said in previous research about the difficulty in measuring teacher effectiveness (Briggs et al., 2014; Dee et al., 2021; Habib, 2017; Ho & Kane, 2013; Jones & Bergin, 2019; Stronge et al., 2011). Jones and Bergin (2019) are most relevant here as they found that principals generally exhibited leniency in evaluations and were biased toward awarding higher scores. The data herein validate Jones and Bergin and questions the practical use of principal-based evaluations for personnel decisions that focus on student performance and school improvement decisions. Considering the possibilities as to why the results are what they are, it is useful to consider the following: Is evaluating teaching that complex and multifaceted (see Garrett & Steinberg, 2015; Walker, 2020) that a five-construct tool is not sufficiently valid and reliable to differentiate teaching performance and effective teaching (Harris & Sass, 2007)? Second, is the training and support provided principals and other evaluators weak and insufficient?

Reliability and Validity of NCEES

In this sample nearly all teachers were evaluated as at least *Proficient*, but not nearly all students are, thus creating a paradox between teacher evaluation ratings and teacher effectiveness in supporting student achievement. As mentioned earlier, this study shows that when evaluating teachers principals rate approximately 97 percent of teachers as *Proficient* or higher, while student performance data shows that schools have mean grade-level proficiency scores (meeting grade level content standard mastery) that range from approximately 50 to 70 percent (NCDPI, 2017). This misalignment may be due to factors related to the instrument's validity and reliability in the evaluation process.

Validity

This study did not attempt to determine the face or construct validity of the NCEES; however, its criterion validity is in question. The criterion validity of the NCEES rubric is questionable when student performance is the criterion to which the evaluation results are compared to determine validity. The standards themselves may not be aligned with student learning outcomes, and in North Carolina, student performance outcomes are not a serious consideration of the NCEES process. In fairness, this may reflect the complexity of teaching, which, as suggested by Muijs and Reynolds (2017), is an activity that may not be conveniently quantified through student learning outcomes. The discrepancy between the high proficiency scores awarded to teachers and the substantially lower student performance scores illustrates this lack of validity or that student achievement as a criterion to evaluate is of minimum value.

Reliability

The high percentage of *Proficient*ly rated teacher scores in this sample and nationally, plus the need for more variation in the teacher ratings, present programmatic concerns worthy of continued investigation. At first glance, the consistently high teacher evaluation scores indicate that the process is highly reliable in North Carolina and nationally. However, the lack of variation in scores is problematic as it fails to show that the instrument and evaluator can reliably discern the difference between effective and deficient teaching. One potential issue may be that evaluator training is insufficient, creating a lack of fidelity in implementation. In many districts in North Carolina, administrators receive initial evaluation training; however, follow-up and continued training do not occur. Once in practice, this lack of continuing support to maintain the reliability of the evaluation process could be a potential source of the highly skewed evaluation outcomes and the poor alignment to student achievement. The data gathered in any evaluation is partly a function of the instrument used and the process used to collect such data, which, therefore, behooves education leaders to assess the validity of the instrument or process used continually.

Training and Support

It is vital to consider that administrator evaluations of teachers could be highly subjective and skewed despite the use of rubrics and standards. Principals may include other factors in determining the evaluation outcomes for teachers, or the instrument used may emphasize teaching aspects that may have little alignment with value-added measures. North Carolina has no definitive standards for how much evidence must be presented to score a particular rating, as the state provides only tersely worded examples. More fully, it could be suggested that what is provided to North Carolina principals and teachers fails to clearly illustrate what constitutes a *Proficient* artifact to show proficiency in the standards. The examples provided arguably lack rigor and completeness, possibly contributing to the absence of variability in evaluation results. The lack of clear and valid exemplars (artifacts) may result in these evaluation expectations being highly dependent upon the perspective of the principal, their skill, knowledge of each of the standards and example artifacts, and their ability to consistently distinguish between what is considered *Developing* versus *Accomplished* or any other combination of evaluation scores.

Theory of Action

In North Carolina, principals must evaluate teachers annually to develop professional growth plans, support personnel actions when needed, and generally report on the quality of teaching (NCEES, 2021). This evaluation process provides both intrinsic and extrinsic motivations. The evaluation process is an extrinsic motivator, where teachers derive some sense of self from how well the principal perceives their teaching ability. Growth plans are a more multifaceted aspect of this process, as previous research has shown that teachers identify with this both through the internal desire to become better and as external motivation imposed by policy (Andra et al., 2015; Taylor, 2023). Beginning in 2019, North Carolina moved away from including the EVAAS when reporting teacher evaluation data and now only utilizes EVAAS measures to report EVAAS aggregated at the school level publicly. EVAAS performance is used for financial bonuses to school teachers and administrators, focusing on external rewards and stating that monetary rewards are sufficient motivators.

Recommendations

It is posited that a review of the annual teacher evaluation in North Carolina should address why the NCEES ratings are highly skewed and have a slight variance. One suggested reason for this outcome is that researchers have offered that effective teaching can be linked to professional learning or growth (Behrstock-Sherratt et al.,2014; Fischer et al., 2018; Muijs et al., 2014;) and may not be easily reflected in value-added contexts and is not connected to evaluation data. In North Carolina, the type of mandated professional learning reflected in teacher professional growth plans largely depends on principal evaluations, EVAAS performance, and teacher input. Suppose the majority of teachers are rated as *Proficient. In that case*, it is logical to ask what the focus of these mandated professional growth plans is, how they align with student outcomes, and if the impact of these growth plans on teacher effectiveness can be determined. The core concern is whether evaluation outcomes drive growth plans. So many teachers are deemed *proficient* or better, so what motivation is there to focus on methods to enhance student academic outcomes or any other aspects of teaching? These questions should inspire further research to understand better how professional growth plans can align with data-driven standards and be linked to measurable outcomes.

Future research may examine the efficacy of principals' training and monitoring when evaluating teachers. Related to this, a review of the instrument to capture teacher behaviors that link to student learning outcomes is warranted. These questions remain unanswered at this point, but if explored, they could improve teacher evaluation in a manner that aligns with student outcomes and supports school improvement.

Third, the context in which schools exist, the ever-present teacher shortage and retention (Irvine, 2019), and the difficulty in removing low-performing teachers may impede teacher evaluations and effectiveness as there are limited options to replace lower-performing educators. Historically and contemporarily, dismissing incompetent and harmful teachers has been difficult as the costs are prohibitive, resulting in significant barriers to improving the educator workforce (Griffith & McDougald, 2016). However, as seen herein, evaluation ratings are not efficient signals of effective teaching. Complicating this analysis is that the NCEES scores are aggregate scores translated into the five rating levels for the entire standard. As such, the ability to gather additional information by analyzing the individual elements that comprise each standard is not possible. To better understand the value of the annual evaluation process, element-level data is suggested to be associated with EVAAS performance, which would shed a more profound light on the process.

Conclusion

Consistent with this research, it is posited that school-level teacher evaluation data needs to provide a reliable and unambiguous representation of teaching consistent with student achievement data, yield growth plans aligned to documented needs, and are a mix of internal and external motivation. Mangiante (2011) was optimistic that the Education Value-Added Assessment System (EVAAS) could point the way to understanding teacher effectiveness and approaches to school improvement. Value-added systems such as EVAAS use test scores to estimate growth expectations; however, it has been suggested that this approach also needs refinement (Condie et al., 2014). For several reasons, value-added measures have been used to assess teacher performance regarding student achievement. For example, it is less sensitive to student race and wealth and that annually, all students can and should improve (SAS, 2012).

What can be done to improve the overall evaluation process and the usefulness of the results? The process is required by North Carolina statute, and it is doubtful to be revised not to mandate these evaluations. Therefore, it is incumbent on school administrators to support high-quality educator evaluations that provide greater discriminatory power with better alignment to student learning outcomes (Staiger & Rockoff, 2010) and teaching behaviors. A more deliberate and concerted effort should be made to train principals to be more discerning when engaging in the evaluation process and to require ongoing support to determine the reliability of the evaluations. The recommended aspects from the MET report may provide a more holistic view of teaching; however, if principals remain the final arbiter rating teacher performance, supporting principals is necessary to maximize the potential of the evaluation process.

References

Andra, C., Liljedahl, P., Di Martino, P., & Rouleau, A. (2015). Teacher tension: important considerations for understanding teachers' actions, intentions, and professional growth needs. In *Proceedings of the 39th Conference of the Psychology of Mathematics Education* (pp. 193-200). PME.

Austin, P. C., & Leckie, G. (2018). The effect of a number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation, 88*(16), 3151-3163.

Bandura, A. (1977). Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, *84*(2), 191–215. https://www.susana.org/_resources/documents/default/3-4524-7-1640787326.pdf

Behrstock-Sherratt, E., Bassett, K., Olson, D., & Jacques, C. (2014). *From good to great: Exemplary teachers share perspectives on increasing teacher effectiveness across the career continuum*. Center on Great Teachers and Leaders.

Bell, B. S., Kanar, A. M., & Kozlowski, S. W. (2008). Current issues and future directions in simulation-based training in North America. *The International Journal of Human Resource Management, 19*(8), 1416-1434.

Berliner, D. C. (1976). Impediments to the study of teacher effectiveness. *Journal of Teacher Education*, *27*(1), 5-13.

Briggs, D. C., Dadey, N., & Kizil, R. C. (2014). *Comparing student growth and teacher observation to principal judgments in the evaluation of teacher effectiveness*. University of Colorado.

Brophy, J. E. (2010). Advances in teacher effectiveness research. *The Journal of Classroom Interaction*, 17–24.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2004). Differentiated teacher effectiveness: Framing the concept. In *Assessing teacher effectiveness: Developing a differentiated model* (pp. 3–11). Routledge.

Center for Education and Policy Research (2013). MET Project. https://cepr.harvard.edu/met-project.

Cheng, Y. C., & Tsui, K. T. (1999). Multimodels of teacher effectiveness: Implications for research. *The Journal of Educational Research, 92*(3), 141–150.

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research, 35(*3), 311-351.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, *40*, 76-92.

Cruickshank, D. R., & Haefele, D. L. (1990). Research-based indicators: Is the glass half-full or half-empty? *Journal of Personnel Evaluation in Education, 4*(1), 33–39.

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching*. Center for American Progress.

Dee, T. S., James, J., & Wyckoff, J. (2021). *Is effective teacher evaluation sustainable? Evidence from District of Columbia Public Schools.* Stanford Center for Education Policy Analysis. https://cepa.stanford.edu/content/effective-teacher-evaluation-sustainable-evidence-dcps

EdNC (2017). *EdExplainer: The ins and outs of EVAAS*. https://www.ednc.org/edexplainer-ins-outs-evaas

Every Student Succeeds Act. (2024). Retrieved from: https://crsreports.congress.gov/product/pdf/R/R45977

Fischer, C., Fishman, B., Dede, C., Eisenkraft, A., Frumin, K., Foster, B., ... & McCoy, A. (2018). Investigating relationships between school context, teacher professional development, teaching practices, and student achievement in response to a nationwide science reform. *Teaching and Teacher Education*, *72*, 107-121.

Gallagher, C., Rabinowitz, S., & Yeagley, P. (2011). *Key considerations when measuring teacher effectiveness: A framework for validating teachers' professional practices* (AACC Report). Assessment and Accountability Comprehensive Center.

Garnett, J. M. (2013). *Teacher effectiveness and experience comparing evaluation ratings and student achievement*. University of Nebraska at Omaha.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224-242.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* National Comprehensive Center for Teacher Quality.

Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, *68*(4), 377-393.

Good, T. L. (1996). Teaching effects and teacher evaluation. In J. P. Sikula, T. J. Buttery, & E. Guyton (Eds.), *Handbook of research on teacher education* (pp. 617–665). Macmillan.

Griffith, D., & McDougald, V. (2016). *Undue process: Why bad teachers in twenty-five diverse districts rarely get fired.* Thomas B. Fordham Institute.

Habib, H. (2017). A study of teacher effectiveness and its importance. *National Journal of Multidisciplinary Research and Development*, *2*(3), 530-532.

Harris, D. N., & Sass, T. R. (2007). *What makes for a good teacher and who can tell?* Paper presented at the 2007 summer workshop of the National Bureau of Economic Research. Cambridge, MA.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* (Research Paper. MET Project). Bill & Melinda Gates Foundation.

Holmes, L., & Schumacker, R. (2020). Latent class analysis of teacher characteristics: Can we identify effective teachers? *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 75-86.

Irvine, J. (2019). Relationship between teaching experience and eeacher effectiveness: Implications for policy decisions. *Journal of Instructional Pedagogies*, *22*. https://files.eric.ed.gov/fulltext/EJ1216895.pdf

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.

Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment*, *24*(2), 91-118.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* (Research Paper. MET Project). Bill & Melinda Gates Foundation.

Lau, F. (2017). Methods for correlational studies. In *Handbook of ehealth evaluation: An evidence-based approach [internet]*. University of Victoria.

Leigh, A. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review*, *29*(3), 480-488.

Mangiante, E. M. S. (2011). Teachers matter: Measures of teacher effectiveness in low-income minority schools. *Educational Assessment, Evaluation and Accountability*, *23*(1), 41-63.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling*. Methodology, 1*(3), 86-92.

Morris, L. S., Grehl, M. M., Rutter, S. B., Mehta, M., & Westwater, M. L. (2022). On what motivates us: a detailed review of intrinsic v. extrinsic motivation. *Psychological medicine*, *52*(10), 1801-1816.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research & Evaluation, 12*(1), 53–74.

Muijs, D., Kyriakides, L., Van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art–teacher effectiveness and professional learning. *School effectiveness and school improvement*, *25*(2), 231-256.

Muijs, D., & Reynolds, D. (2017). *Effective teaching: Evidence and practice*. Sage.

Muñoz, M. A., Prather, J. R., & Stronge, J. H. (2011). Exploring Teacher Effectiveness Using Hierarchical Linear Models: Student-and Classroom-Level Predictors and Cross-Year Stability in Elementary School Reading. *Planning and Changing*, *42*, 241-273.

North Carolina Teacher Evaluation Process (2015). North Carolina Department of Public Instruction.

North Carolina Educator Effectiveness System. (2018). https://www.dpi.nc.gov/educators/home-base/nc-educator-effectiveness-system-ncees

North Carolina Department of Public Instruction. (2018). Accountability Report Archive. https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/school-accountability-and-reporting/accountability-data-sets-and-reports/accountability-report-archive#Tab-2017-18-4723

North Carolina Department of Public Instruction. (2024). EVAAS. https://www.dpi.nc.gov/districts-schools/districts-schools-support/district-human-capital/evaas

Race, T. (2010). Evaluating teacher effectiveness. https://www.wested.org/wp-content/uploads/lpa-evaluating-effectiveness.pdf

Raudenbush, S. W. (2002). Hierarchical linear models: Applications and data analysis methods. *Advanced Quantitative Techniques in the Social Sciences Series/SAGE*.

Rosenbaum, P. R., Rosenbaum, P., & Briskman. (2010). *Design of Observational Studies* (Vol. 10, pp. 978-1). Springer.

Ryan, R. M., & Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. https://digitalwellbeing.org/wp-content/uploads/2020/03/Ryan-and-Deci-2000-Self-Determination-Theory-and-the-Facilitation-of-Intrinsic-Motivation-Social-Development-and-Well-Being.pdf

SAS Institute. (2012). A history of value-added assessment and its research-based implications for policy and practice on teaching effectiveness. https://www.learningforwardpa.org/uploads/2/0/8/5/20859956/a_history_of_value-added_assessment_and_its_research-based_.pdf

Skourdoumbis, A. (2014). Teacher effectiveness: Making the difference to student achievement? *British Journal of Educational Studies*, *62*(2), 111-126.

Slater, H., Davies, N. M., & Burgess, S. (2012). Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics*, *74*(5), 629-645.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, *24*(3), 97-118.

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.

Subedi, B. R., Reese, N., & Powell, R. (2015). Measuring teacher effectiveness through hierarchical linear models: Exploring predictors of student achievement and truancy. *Journal of Education and Training Studies*, *3*(2), 34-43.

Taylor, P. (2023). The complexity of teacher professional growth–unravelling threads of purpose, opportunity and response. *Professional Development in Education*, *49*(1), 16-29.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of educational research*, *68*(2), 202-248. https://mxtsch.people.wm.edu/Scholarship/RER_TeacherEfficacy.pdf

Vosters, K. N., Guarino, C. M., & Wooldridge, J. M. (2018). Understanding and evaluating the SAS® EVAAS® Univariate Response Model (URM) for measuring teacher effectiveness. *Economics of Education Review*, *66*, 191-205.

Vroom, V.H. (1964). Work and motivation. Wiley.

Walker, R. J. (2020). *12 characteristics of an effective teacher*. Lulu Publishing.