Measuring Mathematics and Science Teacher Effectiveness Using Rigorous Course-taking in High School

Elif Oz

eo13b@my.fsu.edu

Florida State University

Paper prepared for the AEFP 44th Annual Conference

Kansas City, MO

March 23, 2019

## Introduction

Teachers are the most important school-level input that can make a difference in students' short-term and long-term outcomes (Goldhaber, 2016; Hanushek, 2016). Hence, evaluating teachers based on their performance and contribution in improving their students' performance is one of the focuses of the accountability policies in the United States. With the expansion of standardized assessments, improvement in students' scores in these tests becomes the main indicator of teacher performance. With the 2002 enactment of the No Child Left Behind (NCLB) Act, followed by the Race to the Top (RTTT), the use of test scores to measure teacher effectiveness may have reached a peak because of its emphasis and encouragement at the federal level (Harris, 2009). With the enactment of Every Students Succeeds Act (ESSA) in 2015, school districts are provided with more flexibility in terms of the student performance measures to be included in their teacher evaluation system by prohibiting states to prescribe the teacher evaluation systems to be used by the districts (Aragon, 2018).

Today, despite the flexibility provided and the prohibition on prescription of teacher evaluation at the state level under ESSA, a majority of states are still using student performance measures in their teacher evaluation policies (Aragon, 2018). For instance, by chaining its statue in 2017, Florida permitted its school districts to develop their own teacher evaluation models, but still required from school districts that the student performance indicators should constitute "at least one-third of" teacher performance measures (Title XLVIII F.S. §1012.34, 2017). This is a promising improvement towards a fairer teacher evaluation policies since using students' test scores as the sole student outcome to measure teacher effectiveness is criticized by researchers and practitioners including teachers who are evaluated based on these measures since it only captures one dimension of teacher effectiveness (Blazar & Kraft, 2017; Croft & Buddin, 2015;

Jackson, 2016; Kane, McCaffrey, Miller, & Staiger, 2013; Newman, Newman, Ridenour, & Morales, 2014; Petek & Pope, 2016). Besides, using student test scores at the high school level is problematic due to the variety of topics covered in different courses and the lack of standardized tests that measure student growth in each topic (Goldhaber, Goldschmidt, & Tseng, 2013; Harris & Anderson, 2013). This makes measuring teacher performance in improving their students' performance harder especially at the high school level.

On the other hand, there exists other non-tested but important outcomes that are found to be associated with students' long-term outcomes. Course-taking in Science, Technology, Engineering, and Mathematics (STEM) is one of these non-tested outcomes that are associated with various short-term and long-term outcomes such as high school graduation, college attendance, college degree attainment, and majoring in STEM fields (citation).

Course-taking is used by some researchers as a long-term outcome to test the persistence of teacher effects on other non-cognitive, non-tested outcomes in the long run (Goldhaber, Theobald, & Fumia, 2018; Jackson, 2012; Petek & Pope, 2016). None of the studies on teacher effects on non-tested outcomes of students used course-taking as a direct, short-term outcome to estimate teacher effectiveness at the high school level.

The purpose of this study to fill this gap in the literature by investigating the existence and magnitude of teacher effects on high school students' course-taking outcomes in mathematics and science and whether the teacher effects on future course-taking are correlated with teacher effects on test scores. The research questions ask:

1. Is there variation across high school mathematics and science teachers in the likelihood of their students' taking higher-level mathematics and science courses in upper grades? Does this association differ by students' demographic background?

2. Do estimates of teachers' value-added based on course-taking and estimates of teachers' value-added based on standardized test scores correlate?

Using administrative data from Broward County Public Schools between the years 2001 and 2015, I build value-added models to estimate teacher effects on the level of the mathematics and science courses students take at upper grade levels. I test the existence of teacher effects using the test that Jackson (2014) recommended and used to test the existence of teacher effects on students' non-cognitive outcomes. I use correlations and simple linear regression to study the relationship between teacher effects on course-taking and teacher effects on test scores.

Results show that there exist teacher effects slightly larger than teacher effects on students' test scores. On average, a one standard deviation difference in teacher effects on the level of mathematics and science courses is associated with 0.12-0.15 standard deviation difference in the level of science course taken later. Moreover, the estimated teacher effects on level of mathematics or science course taken is larger for disadvantaged students such as lower-income students, students from underrepresented minority groups, students with limited English proficiency, or exceptional education students. The results also show weak correlations between teacher effects on level of future science courses taken and teacher effects on test scores. These results are in general in alignment with previous research on teacher effects on non-cognitive outcomes and may indicate that teacher effects on students' future course-taking outcomes in another dimension of teacher effectiveness.

In the following sections, I provide a summary of findings of previous research studies related to teacher effects on students' test based on non-test-based outcomes and some background with the use of teacher VAMs in practice. Then, I introduce my methodology and analytic plan to answer my research questions. Finally, I share my results and conclude by summarizing the findings and their implications including recommendations for future research.

## Literature Review

**Teacher effects on students' test-based outcomes.** Currently, gains in student scores based on standardized tests are the predominantly studied outcome to measure teacher effectiveness for educational researchers and policy makers (Doherty & Jacobs, 2015). It is also a commonly used approach to measure teacher effectiveness in practice due to the studies that report large overall teacher effects on students' test score gains (e.g., Aaronson et al., 2007; Hanushek & Rivkin, 2010; Kane et al., 2013; Kane & Staiger, 2008; Nye et al., 2004). Generally, these studies show that on average, a one standard deviation increase in teacher effects is associated with a 0.11 standard deviation increase in reading achievement and a 0.15 standard deviation increase in mathematics achievement (Hanushek & Rivkin, 2010).

Since it is rarely possible to assign students to teachers randomly in real educational settings, studies that investigate teacher effects on students' test scores use value-added models to provide unbiased estimates of teacher effects without using random assignment. Generally, these models include a variety of different student, classroom, and school characteristics that may affect students' achievement. Empirical and simulation studies that compare teacher effects estimated by value-added models in experimental or quasi-experimental and non-experimental settings show that teacher effects estimated by value-added models under experimental or quasi-experimental settings where students are randomly assigned to classrooms are not different than the teacher effects estimated under non-experimental settings where students are non-randomly assigned to classrooms (Bacher-Hicks, Chin, Kane, & Staiger, 2017; Bacher-Hicks, Kane, & Staiger, 2014; Chetty, Friedman, & Rockoff, 2014; Glazerman, Protik, Teh, Bruch, & Max, 2013; Kane et al., 2013; Kane & Staiger, 2008). Other non-experimental studies that statistically test whether value-added measures of teacher effectiveness are biased due to selection of students to teachers conclude that although bias exists, it is not statistically significant (Glazerman & Protik, 2015; Goldhaber & Chaplin, 2015; Koedel & Betts, 2011).

Although the majority of studies are conducted at the elementary or middle school level, there also exist other non-experimental studies that investigates the existence of teachers' effects on students' test scores at the high school level (Aaronson et al., 2007; Buddin & Zamarro, 2009; Jackson, 2014, 2014; Kane et al., 2013). Both Aaronson et al. (2007) and Buddin and Zamarro (2009) report 0.13-0.14 standard deviation increases in mathematics achievement gains for high-school students as teacher effects increase by one standard deviation. On the other hand, Jackson (2014) found much smaller standard deviation differences in student achievement (confidence interval is between 0.0514 and 0.081 standard deviation for math) for mathematics and English teachers when he accounts for tracking effects. He concludes that students' test scores may not be as useful at the high school level as they are at the elementary school level in determining teacher effects. Moreover, implementation of the value-added models in practice is problematic especially at the high school level due to the variety of courses and lack of standardized tests that are aligned with the content taught (Goldhaber et al., 2013; Harris & Anderson, 2013; Newton, Darling-Hammond, Haertel, & Thomas, 2010). Next, I discuss the use of value-added models of teacher effectiveness in practice and their limitations at the high school level.

***Use of teacher VAMs for teacher evaluation policies.*** Teachers are the most important school-level input that can make a difference in student short-term and long-term outcomes (Goldhaber, 2016; Hanushek, 2016). However, research suggests that teachers' contribution to improved student outcomes is not the same for all teachers (Aaronson et al., 2007; Hanushek &

Rivkin, 2010; Rivkin, Hanushek, & Kain, 2005). Therefore, teacher VAMs have been predominantly used by the educational policy makers at the state and district level to objectively distinguish more effective teachers than their less effective counterparts in schools in the United States (Corcoran, 2010).

Based on these teacher evaluation policies, a considerable portion of the teacher performance measure comes from student achievement determined by the standardized tests. However, if there are no standardized tests for a subject that is being taught by a teacher at a specific grade level, teachers are usually evaluated based on their students' test scores in other subjects that are tested in Florida (Goldhaber et al., 2013). For example, in Broward County, Florida, VAMs of high school teachers who teach non-AP or non-IB science courses other than Biology 1[1] are estimated using students' ELA FSA or general SAT scores because there are no standardized tests administered for these courses (Broward County Public Schools, 2017). Despite the use of ELA scores to capture teacher effects on science achievement in practice, research shows small spillover effect of science teachers on mathematics or ELA achievement (Koedel, 2009; Yuan, 2015). Therefore, using ELA scores assuming that it would capture science teacher effect would yield biased estimates of teacher effectiveness.

Moreover, the pre-test and post-test measures used to calculate high school students' growth over time to estimate teacher VAMs are measuring different knowledge and skills of students. For instance, in Broward County, the pre-test used to calculate student growth in geometry is Algebra 1 EOC scores and for AP courses in all subject areas, FSA ELA scores are

[1] Students who take Biology 1, Biology 1 Honors, Pre-AICE Biology IGCSE Level, Biology Technology, Florida's Pre-IB Biology 1, IB Middle Years Program Biology, Integrated Science 3, and Integrated Science 3 Honors, are subject to take the Biology End-of-Course Assessment. Students' EOC scores are used to estimate VAMs of the teachers who teach these courses.

used as the pre-test score[2] (Broward County Public Schools, 2017). This is problematic since the knowledge and skills measured by the pre- and post-tests varies and using tests that are not designed to measure the same outcome over time to discuss growth is “meaningless” (Raudenbush, 2001, p.508). A comparison of different model specifications for high school teacher VAMs shows that, there is only a weak correlation between the models that estimate teacher effects based on test scores for the same subject and the models that estimate teacher effects based on test scores for a different subject (Goldhaber et al., 2013).

Therefore, using tests that are not designed to measure student outcomes related to the subject matter and using different tests that do not measure the same student outcome as pre- and post-tests to measure teacher effectiveness would result in unfair treatment of high school mathematics and science teachers. Such teacher evaluation practices have brought repercussions at the judicial level wherein teachers who teach non-tested courses filed lawsuits for unfair judgements about their teaching effectiveness (Croft & Buddin, 2015).

In addition to the discussions around the importance of using appropriate standardized test scores to measure student growth and the lack of appropriate test scores at the high school level, relying on only one type of student outcome to measure teacher effectiveness has been criticized by scholars since it only captures one dimension of teacher effectiveness (Blazar & Kraft, 2017; Gershenson, 2016; Jackson, 2016; Petek & Pope, 2016). Research shows that teachers who are highly effective in improving their students’ test scores are not necessarily the ones that also improve other types non-tested but important student outcomes such as high school graduation, college readiness, or attitudes (Blazar & Kraft, 2017; Gershenson, 2016;

[2] Table 1 summarizes the pre- and post-tests used to measure student growth to determine mathematics and science teachers’ VAMs at the secondary level.

Jackson, 2016; Kane et al., 2013; Petek & Pope, 2016). I discuss the results of these studies in the following section.

**Teacher effects on students' non-tested outcomes.** Due to the criticisms about measuring teacher effectiveness based on students' test-based outcomes, a growing body of research has begun to focus on finding ways to measure teacher effects on non-test-based outcomes. These studies consider a range of outcomes from students' affective and behavioral outcomes such as enjoyment, motivation, or suspensions (Blazar & Kraft, 2017; Fleche, 2017; Jackson, 2012; Ruzek, Domina, Conley, Duncan, & Karabenick, 2015) to academic outcomes such as the number of AP courses taken in high school, graduating from high school, or college enrollment (Dahlin, 2016; Goldhaber et al., 2017; Jackson, 2012; Petek & Pope, 2016). Generally, in these studies researchers calculate teacher effects on test-based and different non-test based outcomes and then look for correlations between these two types of teacher effects (Jackson, 2016; Ruzek et al., 2015). They also include teacher effects on non-cognitive outcomes as a covariate in a regression equation in addition to teacher effects on test scores and investigate the change in the explanatory power of the model developed to predict long-term outcomes such as high school graduation, taking the SAT or intend to attend a 4-year college (Blazar & Kraft, 2017; Jackson, 2012, 2016). Their findings generally show low levels of correlation between teacher effects on test-based and non-test-based outcomes and a significant increase in explanatory power when teacher effects on non-test-based outcome are included in the model. These findings are used as an evidence that teacher' effects on non-cognitive outcomes is another dimension of teacher effectiveness (Jackson, 2012, 2016; Petek & Pope, 2016).

Research on teacher effects on students' non-test-based outcomes shows that the variation in these outcomes are associated with who their teachers are. Elementary school teachers' VAMs estimated using students' study habits and social skills as the outcome variable are significantly associated with student suspensions and absences in the following grades as well as students' long-term high school outcomes such as dropping out of high school or SAT score (Petek & Pope, 2016). Another finding is that one standard deviation increase in the estimated teacher effects on students' self-efficacy and enjoyment in mathematics classes increases students' self-efficacy and enjoyment by 0.14 standard deviation and 0.31 standard deviation, respectively (Blazar & Kraft, 2017). The magnitudes of these effects are generally found to be above and beyond teacher effects on students' test-based outcomes (Blazar & Kraft, 2017; Cheng, 2016; Fleche, 2017; Jackson, 2012).

Another non-test based, affective student outcome which is also closely related to future course-taking in the subject matter is motivational orientations (Ruzek et al., 2015). Ruzek et al. (2015) collected data from seventh grade students about their mastery and performance achievement goals in mathematics and use these outcomes to estimate their teachers' value-added. They report 0.08 standard deviation increase in mastery goal orientation, 0.03-0.04 standard deviation increase in the two domains of performance goals of students for a one-standard deviation increase in teacher value-added. Although they found smaller effects of teachers on students' motivation outcomes compared to teacher effects on mathematics achievement (0.12 standard deviation increase for one standard deviation increase in teacher effects), the researchers state that the 0.08 standard deviation increase in mastery goals is not indispensable considering the drastic and steady decrease in student motivation after seventh grade throughout high school years (Ruzek et al., 2015).

Finally, high school course-taking, which is also the focus of this study, is one of the non-test-based outcome included in studies that investigate teacher effects on non-test-based outcomes (Goldhaber et al., 2017; Jackson, 2012; Petek & Pope, 2016b). However, in these studies high-school course-taking is primarily used as a long-term outcome to detect the persistence of teacher effects over time (Jackson, 2012; Petek & Pope, 2016).

Jackson (2012) investigates ninth-grade Algebra 1 and English 1 teacher impact on students' various short-term and long-term "non-test score outcomes" including Advanced Placement (AP) enrollment in North Carolina (p. 2). Using administrative data that includes all ninth-grade students in North Carolina Public Schools between the years 2005 and 2011, Jackson (2012) estimates teacher effects on students' test scores in addition to four short-term, non-tested student outcomes which are namely, GPA, number of absences, suspensions, and on time $10^{th}$ grade enrollment. He used these estimated teacher effects to predict different long-term outcomes, namely dropping out, graduating from high school, the number of AP courses taken in any subject by $12^{th}$ grade, taking SAT, and intending to attend four-year college. He finds that ninth-grade English teacher effects on students' short-term, non-test score outcomes are positively related to the number of AP courses students take throughout high school: a one standard deviation increase in teacher effects is associated with 0.66 standard deviation increase in the number of AP courses taken including school-by year fixed effects and using controls. However, the relationship between Algebra 1 teachers' effect on non-cognitive outcomes and the number of AP courses taken is not significant. He also reports non-significant relationships between mathematics and English teacher effects based on test-scores and AP courses taken. This finding of Jackson (2012) provides limited evidence about direct teacher effects on students' course-taking. He did not calculate direct teacher effects on students' AP course-taking. Moreover, since he is not interested in course-taking in subject matter, he uses AP course-taking in any subject instead of AP course-taking in the subject matter. This may increase the variation unexplained by Algebra 1 teachers' effects on student' non-cognitive outcomes.

In a similar manner, Petek and Pope (2016) investigate elementary school teacher effects on test scores, behaviors (the same outcomes used by Jackson (2012)), and learning skills of students in grades 3-5 and how these effects are associated with long-term outcomes including the number of AP courses taken by 12th grade in Los-Angeles Unified School District. Akin to what Jackson (2012) did, they estimated teacher effects on these short-term outcomes and use these estimated teacher effects to predict long-term outcomes using student data between the years 2003 and 2015. They report significant and positive relationship between teacher effects and number of AP courses taken for VAMs based on test scores and behavior measures but not for VAMs based on learning skills. However, the magnitudes of the associations are small: a one standard deviation increase in teacher test score value-added measure is associated with 0.007 standard deviation increase in the number of AP courses taken and a one standard deviation increase in teacher VAMs based on student behavior measures is associated with 0.014 standard deviation increase in the number of AP courses taken. This study also provides limited evidence for teachers' direct effect on students' course-taking in high school. First of all, the effects estimated are for elementary school teachers, not high school teachers. It is reasonable to expect small associations between elementary school teacher effects on 12th grade outcomes. Students have different teachers before they reach 12th grade and these teachers may also influence students' AP course-taking. Moreover, as in the case of Jackson (2012), the researchers do not estimate direct teacher effects on the number of AP courses taken. Therefore, teachers' direct effects on students' course-taking remains unexplored.

The only study that statistically explores the direct contribution of teachers on their students' course-taking outcomes is conducted by Goldhaber et al. (2017). Instead of using overall teacher effects, the researchers used mathematics and science teachers' licensure test scores to predict advanced mathematics and science course-taking[3] of their students in high school who were in seventh or eighth grade between the years 2010 and 2015 in Washington. They are interested in two different course-taking outcomes of students: taking at least one advanced course in the subject matter and the number of advanced courses they take in each subject. Two licensure test scores of teachers are used as predictors of students' course-taking outcomes: a basic skills test in reading, writing and mathematics and an educator skills test in the subject matter they want to be licensed to teach. The results show that the only significant relationship is between the advanced number of mathematics courses students take and their seventh and eighth grade teachers' basic skills test scores: a one standard deviation increase in teachers' basic skills test scores is associated with .15 standard deviation increase in their students' advanced course-taking in mathematics including school-track-year fixed effects and teacher controls. The relationship is relatively weak and not significant for science course-taking. Taking into consideration the weak association between any teacher characteristics and students' outcomes reported by teacher quality research, this finding is promising despite of the other statistically insignificant results that Goldhaber et al. (2017) report. Exploring overall teacher effects on students' advanced course-taking may provide different results.

Therefore, because of the limited evidence on the accuracy of teacher VAMs to predict mathematics and science teacher effectiveness at the secondary level as well as the criticisms

[3]The mathematics courses that the researchers identified as advanced are trigonometry, statistics, pre-calculus and other mathematics courses that are higher than these three. The science courses that the researchers identified as advanced are chemistry and physics and other science courses that are higher than these two.

about using test-scores as the single important student outcome to measure teacher effectiveness, alternative methods to evaluate these teachers are needed to be studied.

# Methodology

## Research Design

The ideal way to investigate teachers' contributions to their students' future course-taking outcomes is to design randomized experiments where teachers and students are randomly assigned to schools and then compare course-taking patterns of students who are randomly assigned to different teachers. Conducting the study in this way would yield unbiased estimates of teacher effects on students' future course-taking since the characteristics of students that may affect the outcome of interest are randomly distributed across schools and teachers.

Since random assignment of teachers and students to schools and students to teachers cannot be implemented feasibly, simply comparing the course-taking outcomes of students of different teachers would be a naïve estimator and therefore would yield biased results. The reason is that students are sorted to schools and teachers within schools (Aaronson et al., 2007; Bottia, Mickelson, Giersch, Stearns, & Moller, 2017; Kalogrides & Loeb, 2013; Wiswall, Stiefel, Schwartz, & Boccardo, 2014). Certain characteristics of schools and classrooms that they are sorted to may affect their course-taking outcomes (Anderson & Chang, 2011; Bottia et al., 2017; Cha, 2015). For instance, a school that is more STEM-focused (e.g. a STEM magnet) may emphasize the importance of mathematics and science courses more than other schools that are not specifically focused on STEM. Therefore, one would expect the students of this school to take more mathematics and science courses and higher-level mathematics and science courses. Moreover, the students who choose to attend this school would be more motivated to learn mathematics and science than their counterparts who do not go to a STEM school.

In addition to student sorting to schools, students may also be sorted to teachers within schools (Aaronson et al., 2007; Kalogrides & Loeb, 2013). At the high school level, students are generally sorted to teachers based on their ability levels which is also called “tracking” (Spade, Columba, & Vanfossen, 1997). Tracking is known to affect student gains in achievement (Jackson, 2014) and it may also affect the number and level of courses that students take (Kelly & Carbonaro, 2012). The high-achievers may already be motivated to learn mathematics and science, therefore, take higher level courses than their counterparts who are not motivated.

Thus, any naïve estimator that does not account for these confounding factors may result in biased estimates of teacher effects on students’ course-taking outcomes. A more sophisticated way is to build analytic models to predict the outcome of interest by including covariates that are known to be related to the outcome, in addition to teacher effects. Comparisons of teacher effects estimated in randomized and non-randomized settings show that the value-added models under non-randomized settings provide unbiased estimates of teacher effects on students’ test scores (Chetty et al., 2014; Kane et al., 2013; Kane & Staiger, 2008). For this study, I adapt the teacher value-added framework that is used to determine teacher effects on student test-based and non-test-based outcomes to estimate secondary school mathematics and science teacher effects on their students’ future course-taking as a measure of teacher effectiveness (Aaronson et al., 2007; Hanushek & Rivkin, 2010; Henry, Bastian, & Fortner, 2011; Jackson, 2012; Nye et al., 2004; Petek & Pope, 2016). Specifically, I use a two-step value-added model as suggested by Ehlert, Koedel, Parsons, and Podgursky (2014).

In their simplest form in the teacher value-added framework, student scores at the end of the academic year are predicted using students’ previous scores, demographic characteristics, and classroom characteristics that may affect their current achievement which is called a lagged score value-added model (Koedel, Mihayl, & Rockoff, 2015). In a lagged-score value-added model, students' projected scores are calculated based on this model and compared with students' actual scores (Nye et al., 2004). The mean difference between actual scores and projected scores of students of the same teacher would yield the "value" added by that teacher in a specific year. Some models include students' gain scores instead of end-of-year scores as the outcome to account for previous student achievement which is called "gain score models" (Koedel et al., 2015; McCaffrey, 2003). While using a gain score model takes into account measurement errors in students' test scores, it provides poor estimates of teacher effects based on studies that compare different specification models of teacher value-added (Kane & Staiger, 2008; Koedel et al., 2015). Moreover, the lagged-score value-added model is predominantly used in practice by policy makers (Koedel et al., 2015). For these reasons, I choose to employ this type of value-added model of teacher effectiveness.

The first step in a two-step teacher value-added model is to develop a model that predicts the outcome by including the factors that are known to be associated with the outcome of interest. These factors can be at the student, classroom, or school level (in the absence of school fixed effects). Therefore, the general form of the value-added model that I use to predict the likelihood of observing an outcome related to course taking is:

$$Y_{icjst} = \beta_0 + \beta_1 \mathbf{X_{icjst}} + \beta_3 \mathbf{C}_{\boldsymbol{cjst}} + \eta_{st} + e_{icjst} \text{ (1)}$$

where $Y_{ijcst}$ is the either of the outcomes related to the level of course taken (i.e. the level of the next course taken or the highest level of course taken by 12$^{th}$ grade) for student i taught by teacher j in classroom *c* in school *s* in year *t*,

$\boldsymbol{X_{ijcst}}$ is the vector of time-variant and invariant individual characteristics for student *i* taught by teacher *j* in classroom *c* in school *s* in year *t*,

$\boldsymbol{C_{cst}}$ is a vector of classroom characteristics for classroom *c* in school *s* in year *t*,

$\eta_{st}$ is a school by year fixed effect, and

$\mathrm{e}_{\mathrm{icjst}}$ is a student-level residual.

The second step is estimating teacher effects from the student-level residual ($\mathrm{e}_{\mathrm{cjst}}$). The student level residuals have three components: student-level random error term, classroom-level random error term, and teacher effects:

$$e_{icjst} = \theta_j + \varepsilon_{cjst} + \varepsilon_{icjst}$$

Where $e_{icjst}$ is the student-level residual in the model,

$\theta_j$ represents the part of the residual that is attributed to teacher effects,

$\varepsilon_{cjst}$ is a random classroom-level error, and

$\varepsilon_{icjst}$is a random student-level error (Henry et al., 2011; Jackson, 2016; Kane et al., 2013).

The actual student-level and classroom-level residuals are assumed to be randomly distributed therefore, have a mean of zero (Jackson, 2016). Therefore, when the average of the student-level residuals of the same teacher is taken, $\bar{\varepsilon}_{icjst}$ and $\bar{\varepsilon}_{cjst}$ would become zero and the remaining value ($\bar{\theta}_j$) would yield the unbiased teacher effect measure (Ehlert et al., 2014; Jackson, 2016).

As I mentioned before, students can be sorted to schools and teachers within schools. In order to account for student sorting to schools, I include school-fixed effects where students' course-taking outcome is compared within the schools that they are attending. Doing so would account for any school-level differences that can contribute to variation in course-taking outcomes of students.

To account for any potential bias due to student sorting to teachers, I include various covariates that are found to be related to student course-taking. These covariates are summarized in Table 1 and discussed in more detail in the section "Covariates". I also include the level of the current course taken to account for any observable and unobservable covariates that affect the *level* of course-taking.

While sorting bias is a problem in value-added modeling that includes observational data, using value-added models that are complex enough and include multiple years of observations of teachers reduces the severity of the sorting bias based on ability and other unobservable characteristics of students (Koedel & Betts, 2009). For this reason, in addition to using covariates that may affect student sorting to classrooms including previous course levels (Chetty et al., 2014), I use longitudinal data which spans more than 10 academic years and include multiple cohorts of students.

Table 1

*Student- and Classroom-level Covariates Included in the Value-added Models*[a]

| Student Level Covariates | Classroom-Level Covariates |
|---|---|
| Level of the current course taken | Percentage of female students |
| Gender | Percentage of underrepresented minority students |
| Race/ethnicity | Percentage of students who are eligible for free/reduced priced lunch |
| Free/reduced price lunch eligibility status | Percentage of English language learners |
| Limited English language proficiency status | Percentage of exceptional students |
| Exceptional student status | |

[a] School-by year fixed effects are used in all models in addition to the covariates listed.

## Data

The data come from Broward County Public Schools (BCPS) in Florida and include the school years from 2001-2002 to 2014-2015 of grades 9-12. In addition to the transcript data that contains the specific courses students take and the grades they received, the BCPS dataset contains student demographic information including gender, race/ethnicity, socio-economic status, English Language learner and special education status as well as their subject-specific scores from standardized tests.

The dataset also involves the necessary identifiers to link student-level data to the courses they have taken in high school and to the teacher who taught that course in a specific semester which is essential for teacher value-added models (McCaffrey, 2003). School level data including school identifiers are also available in the dataset and can be linked to the student-level data.

**Analytical Sample**

The students who took at least two mathematics or science courses in different academic years between grades 9-12, whose sixth, seventh, or eighth grade mathematics or science FCAT scores are available, and whose mathematics or science teachers can be linked to their data in the BCPS dataset are included in the analysis. There are 85 public or charter schools in the dataset that serve approximately a total of 120,000 students at grade levels 9-12 in each year.

The BCPS dataset spans from 2000-2001 academic year to 2013-2014 academic year. Therefore, the students who were high school students in the Broward County Public School System between these academic years and who satisfy the conditions mentioned above constitutes the sample of the study. Thus, approximately a total of 165,000 unique students from multiple cohorts are included in the analysis.

**Outcome Variables**

Two different outcomes that are related to students' course-taking are included in the analysis. For simplicity, I use the phrase "subject matter" to refer to either mathematics or science. The first course taking outcome is the level of the next course taken in the subject matter in a different academic year. If a student takes two courses in the subject matter in the same academic year, I would consider the highest-level course taken in that academic year. For example, if a student takes a level-three and level-four science course in tenth grade, to determine the effect of his/her ninth-grade science teacher, the value of the outcome 2 would be four since it is higher than three. The second course taking outcome is the level of the highest future course that the students take by $12^{th}$ grade. Since it is not possible for a future teacher to contribute to the past student outcomes, I include only the courses that the student takes after taking that teacher's course. For instance, if a student takes a level-four course in $10^{th}$ grade, a level-five course in $11^{th}$ grade, and another level-four course in $12^{th}$ grade, the value of Outcome 3 for the year that she/he was in tenth grade would be four instead of five since the highest level of course that the students take after grade 10 is four. Following section explains how I determined the level of the courses.

**Determining the levels of courses offered in Broward County.** In order to determine the levels of the mathematics and science courses offered in Broward County, I adapt the technique that Burkam and Lee (2003) employ. I do not use the exact levels that they determined for two main reasons. The first reason is that there are various different courses that are offered in Florida between the years 2000 and 2014 that are not included in the classification of Burkam and Lee (2003). In addition, the transcript data that Burkam and Lee (2003) used were collected in 1992 as part of the National Education Longitudinal Study of 1988 (NELS:88). The types of mathematics and science courses offered, high school graduation requirements, as well as the course-taking patterns of the students have changed since 1992 (Perkins, Kleiner, Roey-Westat, & Brown, 2004; Snyder & Hoffman, 1995, 2002). Since the researchers not only used the rigor of the courses but also used the general sequence that was usually followed by the high school students that they included in their sample, it is likely that the course levels that they determined may have changed due to change in students' course-taking patterns. Therefore, to determine the course levels and course sequences more accurately, I applied their technique to the BCPS dataset but generate my own hierarchy. Whenever needed, I also refer to the course levels determined by the FDOE and the course hierarchy used by the BCPS to help me to assign a level to the course.

## Covariates

Since random assignment of teachers to classrooms is not possible in this study, to account for observable and unobservable factors that can affect course taking in mathematics and science such as interest, motivation, or family background, I include covariates that are found to be related to student course taking in the literature in addition to a lagged score variable and school-by year fixed effects. I will discuss the variables that I include in the model to control for the contribution of these factors to students' course-taking.

**The level of current mathematics/science course taken.** In the teacher value-added framework, students' lagged test scores are included in the model specified in order to take into account the observable and unobservable factors that affect students' end-of-year test scores so that the model provides unbiased estimates of teacher effects (Kane & Staiger, 2008). For this reason, I include the current level of the course as a lagged score in the value-added models that I build.

Although including the level of the current course taken should account for the important factors that affect the level of the future course-taking such as prior interest and motivation in the subject matter, or family involvement, there may still exist some factors that may not be captured by the level of the current course taken. I include covariates that are found to be related to the number and level of courses that students take in mathematics and science in the literature. Next, I describe each of these covariates.

**Gender.** While the gender gap in the number of mathematics and science courses taken in high school is closed, the gap still exists for the highest level of courses. Therefore, I include gender as a covariate in the model. It is a dichotomous variable where it gets the value of "1" when the student is female and gets the value of "0" when the student is male.

**Race/Ethnicity.** The differences in mathematics and science course-taking of students from different racial or ethnic backgrounds is larger than the gender differences. I include this covariate by creating dummy variables, Black, Asian, Latino/a, and other where the reference group is White.

**Free/Reduced price lunch eligibility.** This variable is used as a proxy to the socio-economic status of students' families. It is a dichotomous variable where it gets the value of 1 when the student is eligible for free or reduced-price lunch and gets the value of zero if the student is not eligible.

**Limited English proficiency.** This variable is also a dichotomous variable where it gets the value of 1 if the students' proficiency in English is classified as "limited" and gets the value of zero otherwise.

**Exceptional student education status.** This is a dichotomous variable where it gets the value of 1 if the student needs special accommodations in school and gets the value of zero otherwise.

**Peer effects.** This is a classroom-level characteristic to control for the influence of peers whom the student is taking a mathematics and science course with in a year. I estimate peer-effects by taking the average of the certain characteristics of students who are taking the same course in the same classroom in a certain year. The characteristics are the percentage of female students, the percentage of students who are coming from underrepresented racial or ethnic groups in STEM, the percentage of students who are eligible for free or reduced priced lunch, the percentage of students who are English language learners and the percentage of exceptional students who receive special education.

**Analytic Plan**

In this section, I explain my analytic plan to answer each of my research questions. To answer the Research Question 1, I build two different lagged-score (which is also known as covariate adjustment) value-added models which takes the general form:

$$Y_{icjst} = \beta_0 + \beta_1 \mathbf{X_{icjst}} + \beta_3 \mathbf{C}_{\boldsymbol{cjst}} + \eta_{st} + e_{icjst} \quad (1)$$

where $Y_{ijcst}$ is either of the two course-taking outcomes (i.e. the level of the next course taken or the highest level of course taken by 12$^{th}$ grade) for student i taught by teacher j in classroom *c* in school *s* in year *t,*

$\boldsymbol{X_{ijcst}}$ is the vector of time-variant and invariant individual characteristics for student *i* taught by teacher *j* in classroom *c* in school *s* in year *t,*

$\boldsymbol{C_{cst}}$ is a vector of classroom characteristics for classroom *c* in school *s* in year *t,*

$\eta_{st}$ is a school by year fixed effect, and

$e_{icjst}$ is a student-level residual.

I use ordinary least squares (OLS) to predict students' future course-taking outcomes including dummy variables for each year and school included in the analyses. In each model that I employ, I calculate standard errors that are clustered at the classroom level since the observations at the student level are not independent from each other (Koedel, Parsons, Podgursky, & Ehlert, 2012).

Therefore, to answer Research Question 1, I first run two different models conducted separately for mathematics and science teachers including the one of the course-taking outcomes to each model (see equation (1) for the model). Then, I estimate individual teacher effects on different course-taking outcomes of students by taking the average of the student-level residuals taught by the same teacher. To test the existence of teacher effects on students' course-taking, I randomly pair a teacher's mean residuals for class c with the same teacher's mean residuals for class c'. Then, I regress each classroom mean residual of a certain teacher on the randomly paired mean classroom residual of the same teacher. The logic behind this test is that "if teacher effects exist, the residuals for a teacher in one classroom should be correlated with her/his residuals from another classroom" (Jackson, 2014). Therefore, a significant t-test for the regression coefficient of the matched classroom residual would mean that teacher effects on students' course-taking outcomes exist (Jackson, 2014).

In addition to the raw teacher effect estimates, I also used covariance-based estimation to determine a persistent measure of the variation of teacher effects as used by Jackson (2018). Similar to what I did to test the existence of teacher effects, I calculated mean residuals from equation 1 for each classroom and then estimate the covariance of these mean residuals across classrooms of the same teacher by pairing each classroom with another classroom of the same teacher randomly and replicating the process for 200 times. The square root of the median of the estimated covariances is included as the standard deviation of teacher effects.

As part of the first research question, I ask whether differential teacher effects exist for different groups of students (i.e. females, underrepresented minority students, economically disadvantaged students, English language learners, and exceptional students). I answer this question by estimating teacher effects for these groups of students separately, conducting t-tests to see whether the estimated teacher effects are different for different groups of students.

To answer Research Question 2, I first estimate teacher effects on students' test scores by using the student achievement measures that are used by the BCPS to determine mathematics and science teachers' value-added. Then, I regress these teacher effects on test scores on teacher effects on course-taking to check the level of the association between these two teacher effect estimates.

## Results

### Existence of Science Teacher Effects on Level of Science Course-Taking

To answer the first part of Research Question 1, I look for meaningful variations across high school science teachers in the likelihood of their students' taking higher-level science courses in upper-grades. I estimate two separate teacher effects on the level of course-taking: level of the next science course and the highest level of the science course that students take until they graduate. I will refer to them as teacher effects on next course level and teacher effects on highest course level, respectively when discussing the results.

Table 2 summarizes the teacher effect estimates on the two course-level outcomes and teacher effects estimated based on the test used by the BCPS to calculate teacher VAMs. I report the standard deviation of raw teacher effects as well as the covariance-based estimates of the variability of teacher value-added following Jackson (2018).

Table 2.

*Standard Deviations of Teacher Effects on Different Student Outcomes*

| | Science Score | Next Course Level | Highest Course Level |
|---|---|---|---|
| Raw | 0.22 | 0.24 | 0.19 |
| Covariance-based | 0.11*** | 0.15*** | 0.12*** |

The results show that there exists variation across science teachers in the likelihood of their students' taking higher-level mathematics and science courses in upper grades. The standard deviation estimated based on raw teacher effects on the level of the next science course taken is 0.24 and the covariance-based estimate of the variation in teacher effects is 0.15 standard deviation. This means that on average, one standard deviation increase in teacher effect would be associated with .15 standard deviation increase in the level of the next science course taken. A similar conclusion is also true for teacher effects on the highest science course taken: on average, one standard deviation increase in teacher effects is associated with .12 standard deviation increase in the highest level of the science course taken through 12$^{th}$ grade. When these three teacher effect estimations compared, the largest teacher effect is on the level of the next science course taken.

In addition to the estimation of the variation of teacher effects based on different student outcomes, I also conducted a test for the existence of teacher effects on these outcomes as suggested by Jackson (2014). Based on the significant t-test results for the existence of teacher effects I reject the null hypothesis that teacher effects on the level of the next science course taken and on the highest level of science course taken by 12$^{th}$ grade is equal to zero.

To investigate differential teacher effects on students' course-taking outcomes, I estimated teacher effects separately for students from different demographic backgrounds. In terms of student demographics, I used gender, underrepresented minority status, free and reduced priced lunch status, limited English proficiency status, and exceptional student education status. Table 3 summarizes the variation of teacher effects on the level of science course taken for different student populations. These variations in teacher effects are calculated using the covariance-based estimation method as used by Jackson (2018). Teacher effects on the highest level of science course taken changes based on students' gender, free or reduced priced lunch status, limited English proficiency status, and exceptional student education status always favoring the disadvantaged groups. The largest difference in standard deviation of teacher effects is for limited English proficiency status (SD(LEP)= 0.17, SD(Non-LEP)=0.12). In addition to estimating separate teacher effects for different groups of students, I also ran paired sample t-tests to look for statistically significant mean difference in teacher effects between two groups for each demographic variable. All t-tests provide significant results where the mean teacher effects are always higher for the disadvantaged student group.

Table 3.

*Teacher Effects on Level of Course by Student Background (SD)*

| | Next Course Level (Overall SD = 0.15) | Highest Course Level Overall SD = 0.12 |
|---|---|---|
| Gender | | |
| Female | 0.19 | 0.13 |
| Male | 0.15 | 0.12 |
| Under-represented Minority Status (URM) | | |
| URM | 0.16 | 0.14 |
| Non-URM | 0.17 | 0.14 |
| Free/Reduced Priced Lunch Status (FRL) | | |
| FRL | 0.17 | 0.14 |
| Non-FRL | 0.15 | 0.12 |
| Limited English Proficiency Status (LEP) | | |

| | | |
|---|---|---|
| LEP | 0.17 | 0.17 |
| Non-LEP | 0.15 | 0.12 |
| Exceptional Student Education Status (ESE) | | |
| ESE | 0.17 | 0.14 |
| Non-ESE | 0.15 | 0.12 |

The variation in teacher effects is calculated based on covariance-based estimation method used by Jackson (2018).

**Comparing Teacher Effects on Students' Test Scores and Teacher Effects on Course-levels**

Table 4 summarizes the correlation coefficients that shows the relationship between teacher effects on test scores and teacher effects on the two outcomes for the level of science courses taken. Both correlations are low although significant. Regression analysis results also show a weak relationship where only 5% of the variation in teacher value-added based on test scores is explained by teacher value-added based on course-level outcomes.

Table 4.

*Correlations between Teacher Effects on Students' Course-taking and Teacher Effects on Students' Test Scores*

| | Next Course Level | Highest Course Level |
|---|---|---|
| Science Score | **0.22*** ** | **0.23*** ** |

This finding is in accordance with previous research studies on teacher effects on students' non-test-based outcomes and suggest that teacher effects on level of science course taken in upper grades is capturing a different dimension of teacher effectiveness (Blazar & Kraft, 2017; Jackson, 2018; Petek & Pope, 2016).

## Conclusion

The purpose of this study is to investigate the existence of variation in teacher effects on students' course-taking outcomes in upper grade levels in high school. The results show that there exists variation in teacher effects and this variation is statistically meaningful to claim that teacher effects on the level of the science course taken exist. On average, a one standard deviation difference in science teacher effects is associated with 0.12-0.15 standard deviation difference in the level of science course taken. Similar results are also found for mathematics course levels. This size of variation is close to but higher than the variation of teacher effects on students' test scores. This result is consistent with previous studies that investigated teacher's effects on different non-tested outcomes that shows teacher effects on students' non-tested outcomes are "above and beyond" teachers' effects on test scores (Blazar & Kraft, 2017; Jackson, 2018; Petek & Pope, 2016).

In addition to the promising results for the existence of teacher effects on the level of courses taken in upper grades, I also found larger variations, therefore larger effects of teachers on their students' course-taking outcomes for students who are coming from disadvantaged groups and who are underrepresented in STEM. This may mean that teachers can encourage or discourage their students more when they come from more disadvantaged or unrepresented groups.

High school course-taking in STEM is an important outcome that is strongly associated with college attendance, college degree-attainment, and selection to the STEM fields. Teacher's effects on student's STEM course-taking outcomes can be investigated further by scaling up the population size to more than one districts. Further studies are needed to be conducted to investigate the characteristics of teachers that are associated with higher levels of course-taking in STEM. The literature on students' attitudes towards STEM shows that teacher characteristics such as enthusiasm, expertise in the subject matter, or positive relationship with students is associated with positive attitudes towards STEM field. Such characteristics may also be positively related to students' course-taking outcomes in high school.

## References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achivement in Chicago Public High Schools. *Journal of Labor Eceonomics*, *25*(1), 95–135.

Anderson, R., & Chang, B. (2011). Mathematics Course-taking in rural high schools. *Journal of Research in Rural Education*, *26*(1).

Aragon, S. (2018). *Policy Snapshot: Teacher Evaluations*. Denver, CO: Education Comission of the States. Retrieved from https://www.ecs.org/teacher-evaluations/

Bacher-Hicks, A., Chin, M., Kane, T., & Staiger, D. (2017). *An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys* (No. w23478). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w23478

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles* (Working Paper No. 20657). National Bureau of Economic Research. https://doi.org/10.3386/w20657

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146–170.

Bottia, M. C., Mickelson, R. A., Giersch, J., Stearns, E., & Moller, S. (2017). The role of high school racial composition and opportunities to learn in students' STEM college participation. *Journal of Research in Science Teaching*, *55*(3), 446–473.

Broward County Public Schools. (2017). 2017-2018 Orientation for Broward Instructional Development and Growth Evaluation System (BrIDGES). Broward County Public

Schools. Retrieved from http://www.broward.k12.fl.us/talentdevelopment/news/evaluation/BrIDGES/BRIDGESORIENTATION-TEACH-17_18.pdf

Buddin, R., & Zamarro, G. (2009). *Teacher effectiveness in Urban High Schools* (Working Paper No. WR-693-IES) (pp. 1–42). RAND Education.

Burkam, D. T., & Lee, V. E. (2003). *Mathematics, Foreign Language, and Science Coursetaking and the NELS:88 Transcript Data* (Working Paper No. NCES 2003-01). Washington, D.C.: National Center for Education Statistics.

Cha, S.-H. (2015). Exploring disparities in taking high level math courses in public high schools. *KEDI Journal of Educational Policy; Seoul*, *12*(1), n/a.

Cheng, A. (2016). *Like Teacher, Like Student: Teachers and the Development of Students' Non-cognitive Skills* (Working Paper No. 2015–02). EDRE.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating the bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Corcoran, S. P. (2010). *Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice* (Education Policy for Action Series). New York, NY: Annenberg Institute for School Reform.

Croft, M., & Buddin, R. (2015). Applying value-added methods to teachers in untested grades and subjects. *Journal of Law & Education*, *44*(1).

Dahlin, L. (2016). *The relationship between teaching effectiveness and students' later outcomes in the Pittsburgh Public Schools*. Unpublished manuscript, Cambridge, MA.

Doherty, K. M., & Jacobs, S. (2015). *State of the states 2015: Evaluating teaching, leading, and learning*. Washington, D.C.: National Council on Teacher Quality.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). The sensitivity of value-added estimates to specification adjustements: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy*, *1*(1), 19–27.

Fleche, S. (2017). *Teacher Quality, Test Scores and Non-Cognitive Skills: Evidence from Primary School Teachers in the UK* (CEP Discussion Paper). Centre for Economic Performance, LSE. Retrieved from https://econpapers.repec.org/paper/cepcepdps/dp1472.htm

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125–149.

Glazerman, S., & Protik, A. (2015). *Validating Value-Added Measures OF Teacher Performance* (Working Paper) (p. 32).

Glazerman, S., Protik, A., Teh, B., Bruch, J., & Max, J. (2013). *Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4004*. National Center for Education Evaluation and Regional Assistance. Retrieved from https://eric.ed.gov/?id=ED544269

Goldhaber. (2016). In schools, teacher quality matters most. *Education Next*, *16*(2), 56–62.

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the “Rothstein falsification test”: Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, *8*(1), 8–34.

Goldhaber, D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis*, *35*(2), 220–236.

Goldhaber, D., Gratz, T., & Theobald, R. (2017). What's in a teacher test? Assessing the relationship between teacher licensure test scores and student STEM achievement and course-taking. *Economics of Education Review*, *61*(2017), 112–129.

Goldhaber, D., Theobald, R., & Fumia, D. (2018). *Teacher Quality Gaps and Student Outcomes: Assessing the Association Between Teacher Assignments and Student Math Test Scores and High School Course Taking* (Working Paper No. 185). Washington, D.C.: National Center for Anlysis of Longitudinal Data in Education Research.

Hanushek, E. A. (2016). School human capital and teacher salary policies. *Journal of Professional Capital and Community*, *1*(1), 23–40.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Harris, D. N., & Anderson, A. (2013). *Does value-added work better in elementary than in secondary grades?* (What we know series: Value-added methods and applications) (pp. 1–12). Carnegie Foundation for the Advancement of Teaching.

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *American Educational Research Journal*, *40*(6), 271–280.

Jackson, K. C. (2012). *Non-cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina* (Working Paper No. 18624) (pp. 1–52). Cambridge, MA: National Bureau of Economic Research.

Jackson, K. C. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Eceonomics*, *32*(4), 645–684.

Jackson, K. C. (2016). *What do test scores miss? The importance of teacher effects on non-test score outcomes* (Working Paper No. w22226). Cambridge, MA: National Bureu of Economic Research.

Jackson, K. C. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher*, *42*(6), 304–316.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment* (MET Project Research Paper). Bill and Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

Kelly, S., & Carbonaro, W. (2012). Curriculum tracking and teacher expectations: Evidence from discrepant course taking models. *Social Psychology of Education*, *15*(3), 271–294.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, *28*(6), 682–692.

Koedel, C., & Betts, J. R. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy*, *6*(1), 18–42. https://doi.org/10.1162/EDFP_a_00027

Koedel, C., Mihayl, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economic of Education Review*, *47*, 180–195.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). *Teacher preparation programs and teacher quality: Are there real differences across programs?* (Working Paper No. 79). Washington, D.C.: Center for Analysis of Longitudinal Data in Education Research.

McCaffrey, D. F. (2003). *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: RAND Corporation.

Newman, D., Newman, I., Ridenour, C., & Morales, J. (2014). Value-added methods: Moving from univariate to multivariate criteria. *Journal of Research in Education*, *24*(1), 99–111.

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

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

Perkins, R., Kleiner, B., Roey-Westat, S., & Brown, J. (2004). *The High School Transcript Study A decade of Change in Curricula and Achievement, 1990-2000* (No. NCES 2004-455). Washington, D.C.: National Center for Education Statistics.

Petek, N., & Pope, N. G. (2016). *The multidimensional impact of teachers on students.* (Working Paper). Washington, D.C.: University of Chicago.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*(1), 501–525.

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

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *Journal of Early Adolescence*, *35*(5–6), 852–882.

Snyder, T. D., & Hoffman, C. M. (1995). *Digest of Education Statistics 1995* (Digest of Education Statistics No. NCES 95-029). National Center for Education Statistics.

Snyder, T. D., & Hoffman, C. M. (2002). *Digest of Education Statistics*, *2001* (Digest of Education Statistics No. NCES-2002-130). Washington, D.C.: National Center for Education Statistics. Retrieved from https://files.eric.ed.gov/fulltext/ED467580.pdf

Spade, J. Z., Columba, L., & Vanfossen, B. E. (1997). Tracking in mathematics and science: Courses and course-selection procedures. *Sociology of Education*, *70*(2), 108–127.

Wiswall, M., Stiefel, L., Schwartz, A. E., & Boccardo, J. (2014). Does attending a STEM high school improve student performance? Evidence from New York City. *Economics of Education Review*, *40*, 93–105. https://doi.org/10.1016/j.econedurev.2014.01.005

Yuan, K. (2015). A value-added study of teacher spillover effects across four subjects in middle schools. *Education Policy Analysis Archives*, *23*(30).