Teacher Merit Pay and Student Test Scores: A Meta-Analysis

Lam D. Pham

Tuan D. Nguyen

Matthew G. Springer

Vanderbilt University

April 3, 2017

Author Note

Lam D. Pham, Department of Leadership, Policy, & Organizations, Vanderbilt University; Tuan D. Nguyen, Department of Leadership, Policy, & Organizations, Vanderbilt University; Matthew G. Springer, Department of Leadership, Policy, & Organizations, Vanderbilt University.

We appreciate helpful comments and suggestions from seminar participants at annual meeting of the Association for Education Finance and Policy. We would also like to acknowledge Emily Tanner-Smith and Carolyn Heinrich for their feedback. Any errors remain the sole responsibility of the authors. The views expressed in this paper do not necessarily reflect those of sponsoring agencies or individuals acknowledged.

Correspondence concerning this article should be addressed to Lam Pham, Department of Leadership, Policy, & Organizations, Vanderbilt University, Nashville, TN 37203. Email: lam.d.pham@vanderbilt.edu

Abstract

In recent years, teacher merit pay programs have garnered considerable political and financial support, spurring rapid growth in the number of research studies investigating the association between teacher pay incentives and student test scores. The growing research literature on this topic presents a novel opportunity to synthesize our understanding of merit pay and its influence on student test scores. This study fills that role as a meta-analysis of reported findings from 44 primary studies. Our meta-analysis finds that the presence of a merit pay program is associated with a modest, statistically significant, positive effect on student test scores (0.052 standard deviations). We also find that effect sizes are highly sensitive to program design and study context, which suggests that while some merit pay programs have the potential to improve student test scores in some contexts, researchers and policymakers should pay close attention to how the program is structured and implemented.

Keywords: meta-analysis, review, teacher merit pay, teacher pay for performance, teacher incentive pay

Teacher Merit Pay and Student Test Scores: A Meta-Analysis

During the 2012-13 school year, the most recent year for which national finance data are available, public schools in the United States spent more than $210 billion on salaries, accounting for nearly 60 percent of school expenditures. The primary determinant of teachers' pay is a single salary schedule that standardizes remuneration based on years of experience and highest degree earned. According to the 2011-12 Schools and Staffing Survey, roughly 95 percent of public school districts in the U.S. use a single salary schedule in pay setting. Critics of current teacher salary schedules highlight an incongruence between performance and compensation because years of teaching experience and education level have low correlation with student outcomes (Hanushek, 2003; Podgursky & Springer, 2007, 2010; Springer, 2010). In response to criticisms of input-based salary schedules, more school leaders at the local-, state-, and federal-levels are proposing merit pay systems to better couple teacher performance with compensation.

Teacher merit pay programs, also called incentive pay, performance pay, and pay for performance, offer increased compensation to teachers who meet certain performance criteria, usually involving improved student test scores (Podgursky & Springer, 2007; Springer 2010). Merit pay offers a solution to concerns that compensation based solely on degree attainment and seniority weaken teachers' incentives to exert more effort. Situated in principal-agent theory, pay incentives are designed to help employers motivate workers when individual effort and ability are not easily measured (Dixit, 2002; Heinrich & Marschke, 2010; Holmstrom & Milgrom, 1991). Within education, merit pay offers a theoretically appropriate solution because administrators have limited time and capacity to closely observe teachers' performance in the classroom. Moreover, theories of personnel economics suggest that compensation can: (a) serve as a powerful *motivational* incentive to increase firm performance by encouraging employees to improve their practice and (b) improve the *composition* of the workforce through the attraction and retention of high performers and discouragement of lesser performers from entering or staying in the profession (Lazear 1998; Lazear and Shaw, 2007; Springer, 2010).

In contrast, critics of merit pay contend that assumptions made by principal-agent theory do not hold in teaching. They argue the difficulty inherent in creating a reliable process for identifying effective teachers, measuring a teacher's value-added contribution, eliminating unprofessional preferential treatment during the evaluation process, and standardizing assessment systems across schools (Hatry, Greiner, & Ashford, 1994; Murnane & Cohen, 1986). These criticisms have stigmatized more recent attempts to devise and implement merit pay programs claiming further that teachers do not support merit pay policy (Goldhaber et al., 2005; Podgursky & Springer, 2007) and adding that recent high-profile experiments do not report a statistically significant effect on teacher performance or behavior.

Despite substantial opposition, teacher merit pay programs are growing in popularity with considerable political and financial support. The federal government has awarded over $2 billion in over 100 grants to grantees in more than 30 states and the District of Columbia to design and implement performance pay systems. In Florida, Texas, Colorado, and Minnesota alone, multiple districts have allotted more $550 million to merit pay programs (Springer and Taylor, 2016). Meaningful investigations into the effects of teacher merit pay programs are especially timely in light of increasing financial investments and ongoing controversy over them. Responding to this need, research on merit pay programs has also grown over the last ten years, encompassing both evaluations of programs currently in operation and randomized controlled experiments of discrete interventions.

The growing empirical research on teacher merit pay programs provides a novel opportunity to consider these studies in light of each other. Taking this opportunity, we use a meta-analysis to synthesize the accumulated findings on how merit pay programs influence student test scores. Our meta-analytical approach systematically reviews primary studies in the context of other studies, which has advantages over more traditional narrative reviews of merit pay as in Chamberlin et al. (2002), Harvey-Beavis (2003), Umansky (2005), Podgursky and Springer (2007), Podgursky and Springer (2011), Viscardi (2012), and Imberman and Lovenheim (2015). First, a meta-analysis can investigate whether the effect of merit pay is estimated consistently across the literature. Where the effect is not consistent, a meta-analysis can quantify the extent of variance across studies. Second, a meta-analysis makes use of effect sizes across studies rather than vote-counting methods (i.e., counting *p* values across studies) that overly emphasize statistically significant results. Third, a meta-analysis is especially fitting in light of wide variation in features of merit pay programs. For example, some studies investigate rank-order tournaments where teachers compete for incentives whereas others use fixed performance contracts where teachers receive pay as long as they meet pre-specified performance targets.

With the goal of synthesizing effects from multiple studies, our investigation asks:

1. To what extent does teacher merit pay affect K-12 student test scores?
2. To what extent do results vary across study and program characteristics, such as individual versus group incentives?

Our study is focused on student test score outcomes and the motivational effect hypothesis of performance compensation, as suggested by personnel economic theories. As for compositional effects, we include a brief descriptive summary of the relationship between incentive pay and teacher labor market outcomes such as retention and recruitment because too few primary studies exist on this topic to allow for a more rigorous meta-analytical investigation. Nevertheless, we believe it is important to recognize this second mechanism through which anticipated benefits of merit pay programs may be realized.

The rest of this paper proceeds as follows. In the next section, we explain the context and theoretical framework behind merit pay. Then, we discuss our methodology, including the literature search process, the study inclusion criteria, the coding of primary studies, and the analytic approach. Next, we present results from the main analysis, sensitivity checks, subgroup and moderator analyses, risk of bias assessments, and an analysis of publication bias. We end with a discussion of our findings, their implications, and some preliminary recommendations for future teacher merit pay programs.

Context and Theoretical Framework

History of Compensation in Public Schooling. Teacher compensation in the U.S. traces back to the room and board compensation model common in early 19th century one-room schoolhouses (Protsik, 1995). Under this model, teachers rotated their residence between students' homes and received a small stipend along with room and board. Industrialization in the late 1800s meant greater need for a better educated work force resulting in increased demand for more and better-trained teachers. To meet this need, teacher compensation was reconceptualized to the grade-based model which sought to imitate factory production models by paying teachers based on skill (Podgursky & Springer, 2007). Since it was believed that younger students were easier to educate, secondary teachers were paid more than elementary school teachers (Guthrie, Springer, Rolle, & Houck, 2007). By the beginning of the 20th century, teacher compensation changed again as labor leaders used collective bargaining to advocate for better working conditions and salaries. During this period, the single salary schedule emerged where teachers were paid according to uniform pay steps that meant teachers with the same years of experience and education level were paid equally (Podgursky & Springer, 2007). The single salary pay system continues to be the dominant model of teacher compensation with 95 percent of public school districts using a uniform salary schedule.

Despite its widespread acceptance, the single salary schedule has been criticized because public school administrators could not adjust teacher pay to reflect either performance or labor market realities. One prominent compensation reform model proposed to address this issue is merit-based pay. Merit pay in education is an longstanding idea dating back to Great Britain in the early 1700s (Stucker & Hall, 1971). In the United States, Evenden reported 48 percent of the over 300 cities he studied in 1918 using merit pay of some sort (Evenden, 1919). In the 1920s, when scientific management was common, administrators adapted evaluations from business management to schools, leading to widespread use of merit pay (Johnson & Papay, 2010). These programs quietly dissipated and merit pay faded from public interest between the 1930s and 1950s. Interest in merit pay re-emerged, driven by public concern with international competition during the Cold War. Merit pay programs at the time made use of sophisticated evaluation techniques, such as Teacher Observation Codes meant to aid observers in evaluating teachers (Johnson & Papay, 2010). For all their sophistication, interest in merit pay programs developed during the Cold War waned by the 1970s.

With the release of *A Nation at Risk* in 1983, school districts again revisited merit pay models as alternatives or supplements to the single salary schedule (Podgursky & Springer, 2007). Distinct from knowledge- of skill-based compensation where teachers are paid based on “inputs,” merit-based systems pay teachers, groups of teachers or schools based on outcomes such student test scores, classroom observations, or teacher portfolios. National interest in merit-based pay has steadily grown in recent years with well-known programs such as Denver Public Schools' Professional Compensation System for Teachers (ProComp), Florida's Merit Award Program (MAP), Minnesota's Quality Compensation Program (Q-Comp), Texas' Governor's Educator Excellence Award Programs, and national programs such as the Milken Family Foundation's Teacher Advancement Program (TAP) and the U.S. Department of Education's Teacher Incentive Fund (TIF). While interest in merit pay programs in K-12 public schools has historically waxed and waned, current merit pay programs share at least one common theme with these past movements: theoretical belief in the efficacy of incentives in motivating teachers.

Theoretical Framework. Proponents of merit pay programs often cite principal-agent theory as a framework for understanding the benefits of incentive pay (Burgess & Ratto, 2003; Dixit, 2002). Principal-agent theory predicts that incentives are useful when principals and agents have different goals in the context of information asymmetry (Heinrich & Marschke, 2010). School administrators and teachers pose this principal-agent problem because administrators often do not have the time or capacity to adequately observe what teachers actually do in their classrooms. Also, principal-agent theory is particularly attractive in the context of merit pay in public schools, because the model incorporates organizational structures into an analysis of workers' responsiveness to incentives. In particular, schools improve most when incentives are closely coupled with important student learning outcomes, and teachers are most responsive when they believe their work contributes to the school's goals and when they feel capable of meeting the incentive criteria.

Supporters of merit pay also argue merit pay allows flexibility such that exceptional teachers are compensated for their performance. This argument situates merit pay programs within principles of efficiency where greater effort and better service yield higher salary. Another attractive feature of merit schedules is that employers do not need to specify details for how the outcome is to be achieved (Asch, 2005). In this way, teachers can continue to choose teaching methods they deem best as long as they and their students meet certain target criteria.

In light of these theoretical benefits, scholars in the incentive pay literature suggest two high-level pathways by which incentive pay improves student outcomes, which is also supported in the general personnel economics literature (Podgursky & Springer, 2007). First, merit pay encourages and motivates teachers to improve because their increased efforts will be rewarded (Springer and Taylor, 2016). This pathway is especially well supported by principal-agent theory. Second, pay incentives can be used to attract and retain higher performing teachers who would benefit most from the extra compensation (Ballou & Podgursky, 1998). More studies into this promising pathway have been emerging over the last few years (e.g., Clotfelter, Ladd, & Vigdor, 2011; Dee & Wyckoff, 2015; Fulbeck, 2014; Glazerman et al., 2013; Springer, Swain, & Rodriguez, 2015; Steele, Murnane, & Willett, 2010; Taylor & Springer, 2009), but most of the scholarship on teacher pay incentives has focused on how these programs influence student test scores and emphasized how incentive pay can motivate teachers to improve. In this meta-analysis, we review studies that link merit pay with student test scores through improved teacher motivation and performance, while acknowledging that student test score gains can occur through a combination of improved teacher performance and elevated quality in the teacher workforce.

Balancing voices of support for merit pay, opponents present multiple criticisms of merit pay when applied to teachers, questioning the assumptions of principal-agent theory. First, incentive pay programs assume that organizational goals and criteria for quality performance are clearly defined, assumptions that may not hold for teachers (Dixit, 2002; Lazear, 2001; Mehta, 2013). Teachers and schools have multidimensional educational goals for students such as academic mastery, citizenship, character development, and career preparation, and there is neither agreement on which goals are most important nor are there clear definitions for what makes teachers “good” at achieving these goals (Lazear, 2001). Second, incentive programs assume teachers know how to improve and are simply unmotivated to do so, but researchers have shown that teachers do not always know how to do a better job (Thoonen et al., 2011) and vary in their sense of self-efficacy (Klassen & Chiu, 2011). Finally, critics point out the negative implications of individual merit pay programs on school culture when some teachers receive increased compensation, others do not, and administrators cannot clearly identify why. This argument is especially germane with growing interest in supporting teacher development through collaboration (LaFee, 2003).

The periodic rise and fall of interest in merit pay throughout history along with vigorous debate over its suitability in education suggest a need for careful synthesis of available research findings on teacher merit pay programs. Responding to this need, we use meta-analytical techniques to synthesize the available empirical literature on merit pay in order to provide evidence of its influence on student test scores based on the accumulation of knowledge from multiple studies.

Method

Our study is designed to examine how the presence of a teacher merit pay program affects student test scores and whether results vary across design characteristics of the study or merit pay program. To define the eligibility criteria, literature search, data analysis, and reporting conventions, we follow the Preferred Reporting Items for Systematic Reviews and Meta-Analysis standards as defined by Moher et al. (2009).

Eligibility Criteria

Primary studies eligible for inclusion in this meta-analysis needed to meet the following criteria: (a) the sample is comprised of teachers and students in K-12 education; (b) the teachers are located in a school, district, state, or country with a teacher merit pay program; (c) the study reports quantitative student test scores data such as student performance on math and reading state exams;[1] (d) there is a business-as-usual comparison group[2]; and (e) the study uses a randomized control trial (RCT) or a quasi-experimental design (QED). We also excluded studies that compared the effect of getting a bonus against failing to get a bonus (e.g., Jinnai, 2016).

Literature Search. We obtained primary studies from searching 20 commonly used economic and general social science databases, including ERIC, WorldCat, ProQuest, JSTOR, NBER and EconLit. Through an iterative process, we created the following search string: teach* AND (pay OR incentive* OR salar* OR merit OR "performance pay" OR "pay-for-performance" OR "career ladder"), which returned over 19,000 studies.[3] We also searched for “grey” literature using Dissertation and Thesis Repositories in WorldCat and ProQuest as well as a general Google search for evaluation reports of well-known merit pay programs such as the Teacher Advancement Program (TAP) and the Teacher Incentive Fund (TIF). In addition to searching databases, our literature search also included an examination of reference lists and

[1] We excluded studies that reported proficiency rates or percentage pass as their outcome (e.g., Choi (2015); Dowling, Murphy, & Wang (2007); Ladd (1999) because these outcomes are not comparable to our outcome of interest: average student test scores.

[2] One study (Wellington et al., 2016) falls slightly outside of this scope as the comparison group has three out of four components of a merit pay program without the merit pay component, but it was included due to its national scope. The estimates are robust to the exclusion of this study however. Results are available upon request.

[3] The numbers of returned results for each database are presented in Supplemental Table S1 (online only). We should also note that both Google Scholar and JSTOR limited us to reviewing only 1,000 results each even though more results were returned.

previous reviews of the merit pay literature (Chamberlin, Wragg, Haynes, & Wragg, 2002; Harvey-Beavis, 2003; Imberman, 2015; Podgursky & Springer, 2007; Podgursky & Springer, 2011; Umansky, 2005; Viscardi, 2012).

We limited our focus to publication dates between January 1989 and October 2016. We chose 1989 because of its historical importance in the education accountability movement – the Charlottesville Education Summit where President Bush and most of the nation's governors met to outline education goals that shifted national focus to student outcomes rather than education inputs. We did not limit our search based on language, publication status, or country.

We also included studies conducted outside of the United States (Atkinson et al., 2009; Contreras & Rau, 2012; Glewwe, Ilias, & Kremer, 2010; Lavy, 2002, 2009; Martins, 2009; Mizala & Romaguera, 2005; Muralidharan & Sundararaman, 2011; Santibañez et al., 2007; Woessmann, 2011). Three factors informed this choice. First, almost all investigations in the current literature on teacher merit pay draws on these international studies to inform their methods and framing, and these studies are important voices in the academic conversation over teacher merit pay which should not be ignored when analyzing between-study outcomes. Second, restricting the sample to only studies in the United States would limit the sample unnecessarily, reducing statistical power for meaningful analysis. Finally, many of these international studies were randomized controlled experiments with strong internal validity, making them high quality estimates of the effects of merit pay.

Studies Meeting Eligibility Criteria. Starting with the results returned from our search of databases and previous reviews, we used a three-phase process to screen for primary studies that met all eligibility criteria, as illustrated in Figure 1. First, we read the title, abstract, and introduction for all studies obtained in our original search. We retained a study if the title, abstract or introduction mentioned that the study contained empirical results pertaining to the effect of a merit pay program on student test scores. Some examples of studies excluded in this phase include qualitative reports describing only perceptions of merit pay, investigations that do not mention student test scores as an outcome of interest, theoretical works on the application of merit pay, case studies of fewer than five teachers, studies situated in higher education settings, and multiple reports that mention merit pay without explicit study of its effects. In total, we screened 19,908 records.

[Insert Figure 1 Here]

In phase two, we were left with 137 studies for full text reading where two coders independently assessed whether each study fits the eligibility criteria outlined above. The coders discussed any discrepancies and made exclusion decisions upon consensus or consulted with the third author to resolve any discrepancies. Of the 137, we excluded 93 studies due to lack of relevant student test scores outcomes, non-empirical results, and duplicate reports. For multiple reports from the same study (e.g., a dissertation and corresponding journal article or reports from multiple years for the same evaluation), we kept only the most current publication.

In phase three, we contacted authors to request information when eligible studies were missing key information. We sent e-mails to lead authors requesting information and resent these e-mails if we did not receive a response within three weeks. We excluded eligible studies if key information such as standard errors for effect estimates could neither be calculated nor obtained from the authors. If the standard error or the *t* statistic was not provided, but the significance level was indicated, we used a conservative estimate of the standard error by calculating the *t* statistics for the *p* value corresponding to reported significance levels. Further details on how we calculate standard errors are included in the analysis section below. After screening, we were left with a sample of 44 primary studies that met all eligibility criteria, 33 of which contained effect estimates in mathematics and 27 in reading or English language arts.

Coding Reports

Two of the authors independently coded relevant information for each of the 44 eligible studies using a taxonomy similar to that of Springer and Balch (2010). We describe relevant items in greater detail below. Any discrepancy was resolved by consensus between the two coders and remaining disagreements unresolved by consensus were decided by the third author.

Dependent variable. Our main outcomes of interest were standardized regression coefficients and standard errors from regressing student test scores on an indicator for the presence of merit pay. This index is interpreted as the average standard deviation unit change in students' test scores when their teacher was part of a merit pay program compared to students whose teacher was not part of a merit pay program. To ensure comparability across studies, we recorded standard deviations in the outcome measure in order to standardize regression coefficients when they were not already reported in standardized form. Other coded outcome characteristics include whether student test scores were measured for math, ELA, or a different subject, the instrument used to measure student test scores (e.g., state exams), the unit of analysis, the school level (e.g., elementary, middle, or high school), levels of statistical significance, t-statistics, R^2, sample size, and the covariates in each regression.

Moderating variables. We coded a series of a priori moderators for meta-regression models where we examined how the effects of teacher merit pay varied by study- and program-level characteristics. These moderators were selected based on our reading of the literature and prior work we have conducted on merit pay programs. Specifically, we include the following variables as moderators: (a) whether the study was an RCT; (b) the country where the study took place; (c) whether the merit pay was a bonus or salary bracket increase; (d) whether there was professional development available for the treatment group in addition to merit pay; (e) whether there was a group incentive at the teacher-team or school level; (f) how long the merit pay program was implemented; and (g) whether the study was peer-reviewed. We also coded other study characteristics such as: (a) the identification strategy; (b) whether it was an evaluation of an existing intervention; and (c) whether pre-treatment equivalence was established between treatment and comparison groups. Finally, we coded characteristics of the merit pay program studied in each report including: (a) criteria for receiving an merit pay award; (b) the minimum, maximum, and average amount of the merit pay received; (c) whether merit pay was one component of a larger program that includes additional interventions such as teacher training; and (d) whether the teachers receiving a merit pay award also received recognition in the form of a public announcement.[4]

Analytic Strategy

Analysis of these data follow methods as presented by Borenstein, Hedges, Higgins, & Rothstein (2009). Since most studies reported effect sizes at multiple time points, with multiple estimation techniques, for different subject areas, and at different levels of analysis, our synthesis of 44 studies contains 287 effect sizes. Below, we describe analytical decisions in selecting models, accounting for these multiple within-study outcomes, reconciling studies that use similar data, and assessing risk of bias from differences in study quality.

One important choice for this meta-analysis was the decision between a fixed-effect versus a random-effects model. The fixed-effect model assumes a common true effect size across all studies, whereas the random-effects model allows the true effect size to vary across studies

[4] All coded variables and their descriptions can be found in Supplemental Table S2 (online only).

(Borenstein, et al, 2009). Mechanically, the fixed-effect model assigns weights (W_i) to each study (i) using the inverse of each within-study variance (V_{y_i}):

$$W_{i,Fixed} = \frac{1}{V_{y_i}} \tag{1}$$

In contrast, the random-effects model weights studies using both the within-study variance and the estimated between-study variance (T^2):

$$W_{i,Random} = \frac{1}{V_{y_i} + T^2} \tag{2}$$

For this investigation, a random-effects model is most fitting because substantial variation exists across studies in terms of intervention characteristics such as the amount of incentive pay offered, how long the programs were implemented, and the criteria teachers must meet in order to receive the incentive. Moreover, we do not expect the effect of teacher merit pay programs to be homogenous across different populations and settings. Below, we also present quantitative evidence that a random-effects model is more appropriate than the fixed-effect model.

In order to account for multiple outcomes and time points within a study, we chose not to treat each within-study outcome as separate, because this method unfairly assigns more weight to studies with more reported outcomes, and it assumes within-study outcomes are independent. For example, math and reading scores within a study of teachers in the same district receiving similar pay incentives will have some amount of correlation. To account for multiple within-study outcomes, we computed the mean of all outcomes within each study and used this average as the unit of analysis. To calculate standard errors for each study, we used the variance formula presented by Borenstein et al. (2009), which has the advantage of taking covariance between different outcomes (i, j) into account:

$$var\left(\frac{1}{m}\Sigma_{i=1}^{m}Y_i\right)=\left(\frac{1}{m}\right)^2\left(\Sigma_{i=1}^{m}V_i+\Sigma_{i\neq j}r_{ij}\sqrt{V_i}\sqrt{V_j}\right) \qquad (3)$$

One drawback to this variance formula is that it requires correlations between each outcome, r_{ij}, a measure rarely reported. Without access to this information, we estimate a correlation of 0.5 between each outcome as a median measure between $r_{ij}=0$, which will certainly underestimate the variance, and $r_{ij}=1$, which will overestimate the variance. We also check this variance at different correlations below. Our results are fairly robust across a wide range of r_{ij}.

An alternative method is to use a robust variance estimation to account for the non-independence of effect sizes (Hedges et al., 2010). This method adjusts the standard errors to account for the shared variance due to the study-level characteristics for a given value of ρ, the expected correlation among the dependent effects. Following Tanner-Smith and Tipton (2013), we tested values of ρ ranging from 0 to 0.9 in increments of 0.1. The results using robust variance estimation are similar across different values of ρ and the point estimate and standard error is the same as the our estimate with $r_{ij}=0.5$ to the third decimal place. However, a drawback of robust variance estimation is that traditional measures of heterogeneity such as I^2 or Q are not available for analysis. Consequently, we use the variance formula presented in Borenstein et al. (2009) as the main analytical technique.

Our search strategy sometimes yielded multiple policy reports and research articles studying the same merit pay program. To avoid overweighting results from almost identical data, we only kept the most recent results if multiple versions of a study were published by the same authors. If different groups of researchers investigated the same merit pay program and its effects on the same sample of students in overlapping years, we averaged their results, giving each study equal weight. Following this method, we averaged together four primary studies utilizing data from the School-wide Performance Bonus Program (SPBP) in New York City Public Schools.[5] Also, two reports evaluated the Teacher Advancement Program (TAP) in the same district with overlapping time periods and were also averaged together.[6] After combining these reports, we were left with a final analysis sample of 40 studies.[7]

Risk of Bias Analysis

The process of selecting studies for a meta-analysis presents a number of challenges, with competing schools of thought on the optimal approach. The selection process is important because inclusion or exclusion of studies determines the scope and validity of meta-analytic results. We opted to use an inclusive approach which may make the comparison and synthesis of studies questionable given that studies included in the analysis are decidedly different and poorly produced studies may inject considerable bias. To address this concern, we consider two separate approaches: the *critical evaluation approach* as defined by Lam and Kennedy (2005) and the *quality rating approach* as defined by Lipsey and Wilson (2001).

In the critical evaluation approach, each study included in our review was given a score from 0 to 15 based on the number of minimum quality criteria met. Table 1 presents the fifteen quality criteria. We included only studies with a threshold score of 13 or higher out of 15 and then compared whether our summary effect including all studies differed from our estimate when we included only studies meeting the minimum quality threshold.

[Insert Table 1 Here]

[5] The four reports/articles on SPBP include Fryer (2013), Goodman & Turner (2011) Marsh et al., (2011), and Springer & Winters (2009).

[6] The two TAP reports include Schacter & Thum (2005) and Schacter, Thum, Reifsneider, & Schiff (2004).

[7] The results remain qualitatively and quantitatively similar if we treat these reports as individual studies. For the remainder of this paper, we refer to 40 "studies" as our primary analysis sample, recognizing that one such "study" is an average of four reports or articles on SPBP and another is an average of two reports on TAP.

In the quality rating approach, each of the study authors independently rated each study holistically using our professional judgment of the quality of the study on a scale of 1 to 5 where 1 has high risk of bias and 5 has low risk of bias. Table 1 also presents some criteria we considered when determining our ratings. After independently rating each study, we discussed our individual scores until we obtained consensus on a final quality rating for each study. We note that the vast majority of our independent quality ratings were either exactly the same or differed by one point. There were two cases where the independent ratings differed by two points for two observers, but differences were not due to disagreements about quality of analysis; rather, they were due to differences in how strongly we believed the primary studies' argument that their sample of teachers working under merit pay programs were comparable to the comparison group teachers. Our individual ratings never varied by more than two points.

Results

Our goal was twofold. First, we analyzed the extent to which teacher merit pay affects student test scores. Second, we investigated how the results vary across both study- and program-level characteristics. In this section, we present our findings on how merit pay is associated with student test scores, how results vary across study and program characteristics, and whether the effect estimates are robust to different decision rules. We conclude with a qualitative presentation of what previous studies have concluded about the effects of merit pay on the recruitment and retention of teachers.

Table 2 presents descriptive information about primary studies included in the analysis. Effect sizes ranged from -0.366 to 0.690 with a mean value of 0.084. Approximately 74 percent of the effect sizes recorded in our review are positive, with one study reporting a significant negative effect (Martins, 2009).[8]

[Insert Table 2 Here]

In terms of study characteristics, almost half of the studies are peer-reviewed publications, where peer-review is defined as the peer-review journal publication process or a peer-review process at a large research firm. Twenty-five percent of the studies in our sample used a randomized control study design, most of which were published post-2005. Sample sizes ranged from 323 to 8,561,194 students or 92 to 43,251 schools with an average sample size of approximately 594,751 student or 8,254 school observations.

The merit pay treatment duration ranged between one and twelve years with an average treatment length of approximately four years. The type of awards received by teachers included gifts, one-time bonuses, and permanent salary increases, ranging in value from approximately $26 to $20,000 US. The smallest award amounts are from merit pay programs implemented in developing countries or those programs that offered gifts as awards (Glewwe et al., 2010). About one in five of the primary studies evaluated a merit pay program with a job-training component, a feature that has become a requirement of the federal Teacher Incentive Fund program.

Effect of Merit Pay on Student Test scores

Table 3 presents random-effects estimates of the association between the presence of a teacher merit pay program and student test scores for all studies meeting our inclusion criteria. As reported in Panel B, the overall effect estimates indicate that, on average, teacher participation in a merit pay program was associated with a 0.052 standard deviation increase in student test scores and a fairly precise standard error of 0.008, with lower and upper bound

[8] Martins explains the negative finding by stating that the results “are consistent with incentives-related disruption in collaborative work in schools” (p. 17), suggesting that pay incentives led to decreased student test scores because the competition for bonuses had adverse effects on collegial support among teachers.

estimates of 0.037 and 0.068, respectively. Based on empirical benchmarks established by Hill and colleagues (2008), an effect size of 0.052 is roughly equivalent to 4 additional weeks of learning assuming a standard deviation of 0.40 per year and 36 weeks in a school year. When we subset our analysis to only studies conducted in the United States, the summary effect estimate decreases to 0.035, or about 3 additional weeks of learning, but remains significant. While we believe the strong internal validity of international studies warrant their inclusion, the subset of studies conducted in the U.S. contains less variation in economic and social contexts, and we continue to find evidence of significant effects. The smaller summary effect size suggests that merit pay may have a smaller influence in U.S. schools compared to other countries and illustrates how implementation of merit pay can vary depending on school context.

[Insert Table 3 Here]

We also produce estimates by two subject areas most reported in the literature, mathematics and reading/English language arts (ELA). We find that, on average, the influence of merit pay on student test scores is relatively similar across the two subject areas. The average effect for math and ELA test score outcomes are 0.066 and 0.037, respectively. These estimates are not statistically different from the overall effect.

The random-effects model is the most conceptually appropriate model due to substantial variations across studies in terms of study- and program-level characteristics. However, we also present empirical evidence that a random-effects model is more appropriate than a fixed-effect model. Panel B of Table 2 includes three statistics relevant to the heterogeneity of study effects for both the main effect and by subject. The proportion of observed variance that reflects true heterogeneity in effect sizes (I^2) is 89.564, indicating that less than 11 percent of the total variation can be attributed to random error. Cochrane's *Q* statistic, which is a classical measure of heterogeneity used in the meta-analytic literature, tests the null hypothesis of homogeneity across studies. With a p_Q less than .001, we find evidence to reject the null hypothesis that the true dispersion of effect sizes is zero. In other words, effect sizes are heterogeneous across studies. Relatedly, the estimated variance of the distribution of true effect size parameters across studies (T^2) is 0.001, suggesting a tight distribution of effect sizes across studies. Together, these three measures suggest that there is heterogeneity in effect sizes, justifying the random effects model; however, the effects from different studies are not widely dispersed.

Moderators of the Effect of Merit Pay on Student

While our results indicate that merit pay has a modest, statistically significant effect on student test scores, these estimates vary depending on the context and implementation of the incentive program. To illustrate, Figure 2 presents a forest plot of the overall random-effects model. Each row represents a primary study in our meta-analysis, plotted according to the standardized effect estimates, or roughly the difference between the average score of students enrolled in a classroom taught by a teacher eligible for merit pay and the average score of students in the control or comparison condition used by the study authors after controlling for various student, teacher, and school characteristics. The dotted vertical line intersecting with the black diamond at the bottom of the graph shows the average effect size across all 40 studies, assuming a correlation of 0.05 between multiple within-study outcomes.

[Insert Figure 2 Here]

We explore how these results vary across both study- and program-level characteristics using a number of potential moderators of the effect of merit pay on student test scores. Study-level characteristics include whether the study was peer-reviewed or randomly assigned units to a control or treatment condition. Program-level characteristics include whether the incentive structure was a rank-order tournament, rewarded teachers at the group level, or included merit pay plus some other reform component such as job training. As displayed in Table 4, we find that peer-reviewed studies and studies employing a randomized research design have slightly larger effects than those reported for the complete sample. In terms of program-level characteristics, we find that incentive pay programs employing a group incentive design produce an effect over two times the average study in our sample (0.111 vs. 0.052), which lends support to the shared nature of teaching and learning hypothesis. Interestingly, the studies conducted on merit pay programs that include an on-the-job training component show no statistically significant difference in effect.

[Insert Table 4 Here]

Sensitivity and Robustness Checks

We examine the robustness of our findings to a number of common threats identified in the meta-analytic literature, including publication bias, risk of bias, non-independence of effect sizes, and the unit of analysis used in the evaluation of the merit pay program.

Publication Bias. A common threat in the meta-analytic literature is publication bias. That is, the literature included in our study may be systematically unrepresentative of the true population of completed studies on merit pay. To explore this threat, Figure 3 presents a contour-enhanced funnel plot, which is designed to aid in differentiating asymmetry due to publication bias from that due to other factors. The contour overlay shows if studies appear to be missing in areas of statistical non-significance (the area inside the inner most funnel) or in the areas of higher statistical significance (the area outside the inner funnel). If studies are missing in areas of statistical non-significance, this adds credence to the possibility that the asymmetry is due to publication bias. Studies missing in the area of high statistical significance suggest the cause of the asymmetry may be more likely due to factors other than publication bias, such as variable study quality. Asymmetry to the left or right of the center indicates that studies are systematically more likely to have found either negative or positive results, respectively.

[Insert Figure 3 Here]

Figure 3 provides no evidence that our analysis lacks small studies with small effect sizes. However, there is asymmetry due to a lack of negative effect sizes in studies with "large N". In other words, our large N studies tended to have positive effect sizes, showing some evidence that studies with significant results are more likely to be reported.

While these tests suggest a possible bias where null results are not published, we argue that the extent of bias is not large for the sample of studies we wish to investigate. First, there is a high concentration of studies around the zero effect estimate that are precisely measured. Second, we adopted an exhaustive and relatively inclusive literature search process. Meta-analyses often exclude non-journal publications such as research reports and program evaluations. The presence of these types of studies in our analytic sample helps rule out the possibility of insignificant or unfavorable results being excluded from our analytic sample. Third, Egger's test for asymmetry of the funnel plot indicates there is no bias from smaller studies with small effects.[9] Rather, the asymmetry is coming from studies with larger sample sizes and modestly larger effect estimates.

Moreover, most studies included in this review have relatively large sample sizes with about 70 percent containing sample sizes larger than 10,000 students. We argue that these types of large-scale studies are likely to be published even if they had found null effects, because they have important implications for researchers and policy-makers thinking about merit pay

[9] We selected the Egger et al. approach over other strategies given that the Begg method has very low power to detect biases (Sterne, Gayaghan, & Egger, 2000) and we have a number of imbalances in control and treatment sample sizes due to the non-experimental nature of many studies included in our review.

programs. Indeed, several large randomized control trials did publish null results (e.g., Fryer, 2013; Springer et al., 2011). Any concern that effect sizes are larger in smaller studies is driven mostly by one study (Atkinson et al., 2009) with a larger effect (0.690) and relatively small sample size (181), but this study is given very little weight in our model. Consequently, we do not believe publication bias is a serious threat to the findings reported.

Risk of bias. Table 5 presents results from our risk of bias analysis using both the critical evaluation approach and the quality rating approach. Our critical evaluation approach rated each study from zero to fifteen based on the number of quality criteria met. Studies meeting our inclusion criteria received a rating of thirteen or above. In total, 36 of 40 studies met the inclusion criteria and, as displayed in Table 5, our point estimates are very similar to the estimates for the complete sample (0.043 and 0.052, respectively), suggesting the risk of bias from study selection criteria is low. Using the quality rating approach, we assigned each study a rating of one (high risk of bias) to five (low risk of bias) and included only studies with a rating of four or above. As displayed in Table 5, 18 of 40 studies received a quality rating of four or five, and the results indicate that our findings do not change substantially when we drop the high risk of bias studies from our analysis.

[Insert Table 5 Here]

Unit of Analysis. Table 5 also presents summary effects based on the unit of analysis, either at the school-level or the student-level. There are 9 studies at the school-level, 29 studies at the student-level, and 3 studies at the teacher/grade-level.[10] A concern in pooling these studies together is that there may be large differences between the summary effects for the different units of analysis. The results in Table 5 show that the summary effects at the school- and student-

[10] When the four reports on the SPBP experiment were averaged together, the aggregated group of studies contained both student-level and school-level outcomes, resulting in 41 effect estimates for this analysis.

level are both positive and statistically significant. The summary effect at the school level is larger, at 0.066 compared to 0.051 at the student level, and there is substantial overlap in the 95 percent confidence intervals of the estimates at the school- and student-level. Consequently, we pool these studies together instead of doing separate analyses where the limited sample size would lead to imprecise estimates.

Non-independence of effect sizes. Another threat to the validity of meta-analytic results is the non-independence of effect sizes within studies (recall that our analysis relies on a total of 287 within-study effect estimates). To address this concern, we estimate the overall summary effect for a range of correlations between multiple within-study outcomes, r_{ij}. Figure 4a shows the summary effect ranges from 0.069 to 0.046 for r_{ij} of 0 and 1, respectively, at intervals of 0.1. Additionally, as displayed in Figure 4b, when robust variance estimation is used, the summary effect is tightly bounded around 0.051 for a range of ρ ranging from 0 to 0.9 at intervals of 0.1. Altogether, using an r_{ij} of 0.5 in estimating the summary effect appears reasonable, and results are robust across different values of r_{ij}.

[Insert Figure 4 Here]

Summary of Literature on the Effect of Merit Pay on Teacher Retention

Theories of personnel economics suggest that merit pay can improve the composition of the workforce through the attraction and retention of high performing teachers. While this is a primary mechanism through which merit pay programs may realize their intended purpose of improving student outcomes and learning opportunities, the vast majority of primary studies included in our analysis did not focus on teacher attrition. This is problematic because teacher attrition is an important outcome for districts and policy makers; and if teacher attrition is high, the effect of merit pay on student test scores may be driven by the effectiveness of the teachers who remain relative to the effectiveness of new teachers or substitute teachers replacing the teachers who left. High teacher attrition due to merit pay programs could then have adverse unintended consequences for schools and districts because there is high cost associated with recruiting and training new teachers.

The limited number of primary studies on how merit pay teacher is associated with teacher turnover precludes a full meta-analysis, but we gathered fifteen studies investigating teacher labor market outcomes. Table 6 shows that six of these studies found mostly significant positive effects on teacher retention or recruitment (Booker & Glazerman, 2009; Clotfelter, Glennie, Ladd, & Vigdor, 2008; Cowan & Goldhaber, 2015; Fulbeck, 2014; Glazerman et al., 2013; Springer et al., 2010). Seven of these studies show some positive results but the findings were inconsistent (Choi, 2015; Glazerman & Seifullah, 2012; Hough, 2012; Springer, Lewis, et al., 2009; Springer, Podgursky, et al., 2009; Springer et al., 2015; Steele et al., 2010). Finally, two studies found mostly insignificant effects (Dee & Wyckoff, 2015; Fryer, 2013).

[Insert Table 6 Here]

While we are wary of relying on these vote-counting methods to summarize significant study effects, these studies offer preliminary signs that merit pay programs have the potential to decrease overall teacher turnover and increase recruitment to high poverty schools. Many of these studies found positive effects at least while the merit pay program was in operation and suggest that positive results are most consistent among teachers who are actually eligible to receive the award. For example, Cowan and Goldhaber (2015) found that the Challenging Schools Bonus increased the proportion of targeted National Board Certified teachers in high poverty schools, but overall turnover rates remain unchanged. It is unclear whether the effects of merit pay in these studies persist over time with some researchers finding significant effects only in the first few years of program implementation (Glazerman & Seifullah, 2012; Springer, Lewis, et al., 2009) and others finding effects only when schools have implemented merit pay for several years (Choi, 2015). Many questions remain, such as whether merit pay can attract and retain relatively more effective teachers, and we urge investigators to further explore how merit pay affects teacher recruitment and retention.

Discussion

Our meta-analysis investigates primary studies that report on the effect of merit pay on student test scores. Overall, we find a modest, statistically significant positive association (0.052 standard deviations) between teacher merit pay programs and student test scores. In substantive terms, the effect is roughly equivalent to 4 additional weeks of learning.

Theoretically, merit pay has the potential to improve student test scores by either motivating teachers to improve their performance or by attracting and retaining more effective teachers. However, several previous large-scale empirical studies, especially randomized controlled experiments, which examined the effects of merit pay have resulted in null effects on student test scores (e.g., Fryer, 2013; Springer et al., 2011; Springer et al., 2012). Our review broadens this perspective by including studies conducted outside the United States, reports evaluating programs where merit pay is part of a larger intervention that may include teacher training, and unpublished dissertations and theses. We acknowledge that our inclusive approach may have resulted in the inclusion of less methodologically rigorous studies, but our results are robust to the exclusion of lower quality studies. Indeed, we find that aggregating results from multiple studies across different cultural, economic, and political contexts suggests that incentive pay, as argued by merit pay advocates, offers a promising strategy to improve student test scores.

We find that the effects of merit pay programs are sensitive to how the program has been designed and implemented. Therefore, the more pertinent question may be how merit pay programs should be designed if positive effects are detected in some contexts while null or even negative effects are detected in others. Our evidence, for example, suggests that group incentives result in larger positive effects on average than incentives given to individuals. Numerous other design questions deserve further exploration. Incentives accompanied by school-wide public announcements (Glewwe et al., 2010), incentives making use of loss aversion (Fryer, Levitt, List, & Sadoff, 2012), and incentives awarded based on sophisticated composite evaluative criteria (Dee & Wyckoff, 2015) have been explored by different researchers and shown to have positive effects. We advise continued study into whether the effects of these various program features vary across different contexts as incentive pay program design can take on a number of different forms with differing tradeoffs (Barlevy & Neal, 2011; Neal, 2010; Springer & Balch, 2010; Springer, 2012; Ritter & Barnett, 2013).

Other relevant program features which we lacked the information to investigate more deeply include the amount of pay, how long the program has been implemented, and whether school staff are well informed about the program's guidelines. The latter proved salient in the recent national impact evaluation of the Teacher Incentive Fund where approximately 40 percent of treatment teachers were unaware they were eligible for a bonus (Wellington et al., 2016). We suggest more explicit attention to these types of program features in future research in order to further elucidate how different program features could result in different outcomes.

Finally, this meta-analysis contributes another step toward understanding the different motivational aspects of compensation as grounded in the general personnel economics literature. Primary studies in our sample often suggest that merit pay encourages teachers to increase their effort, a pathway well supported by principal-agent theory. Indeed, when aggregating these studies together, our evidence supports the notion that opportunities to earn pay incentives can lead to improved test scores, perhaps through some increased teacher effort (or, nefariously, gaming of the performance measure system). We note that our study is specifically focused on incentivized outcomes: student test scores. While a sensible goal of any incentive program should be to bring about direct improvements to the targeted outcome, increased teacher effort could feasibly improve other student outcomes (e.g., alternative tests, attendance, students' self-confidence). Future evaluations of merit pay programs should pay closer attention to these alternative outcomes.

Teacher recruitment and retention, however, is another theoretically supported pathway through which merit pay can affect student test scores. Our qualitative review of the emerging literature on this pathway suggests that the positive effect reported in our primary studies may partly be the result of lower levels of teacher turnover. Certainly, some studies find that pay incentives have the potential to increase teacher recruitment to high-need schools, and decrease attrition. We highly suggest continued investigation into teacher labor market outcomes, especially into the effects of pay incentives on the mobility patterns of highly effective teachers and the exit decisions of traditionally low-performing teachers. Regardless of the outcome, our study exposes the sensitivity of effect sizes to program design and study context, and we urge researchers and policy-makers to pay careful attention to these features when evaluating the effectiveness of incentive pay programs.

References

Asch, B. J. (2005). The economic complexities of incentive reforms. *High-Performance Government: Structure, Leadership, Incentives, Santa Monica, Calif.: RAND Corporation, MG-256-PRGS*, 309–342.

*Alafita, T. A. (2003). The effects of performance pay for teachers: An analysis of Arizona's Career Ladder program. *Unpublished Doctoral Dissertation, The George Washington University*.

*Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., & Wilson, D. (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics*, *16*(3), 251–261.

*Balch, R., & Springer, M. G. (2015). Performance pay, test scores, and student learning objectives. *Economics of Education Review*, *44*, 114–125.

Ballou, D., & Podgursky, M. (1998). Teacher recruitment and retention in public and private schools. *Journal of Policy Analysis and Management*, *17*(3), 393–417.

Barlevy, G. and Neal, D. (2011). Pay for Percentile. NBER Working Paper No. 17194. Cambridge, MA.

*Barrera-Osorio, F., & Raju, D. (2015). Teacher performance pay: Experimental evidence from Pakistan. *World Bank Policy Research Working Paper*, (7307).

*Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: results from a social experiment in Mexican high schools. *Journal of Political Economy, 123*(2), 325–364.

Booker, K., & Glazerman, S. (2009). Effects of the Missouri Career Ladder Program on Teacher Mobility. *Mathematica Policy Research, Inc.*

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Wiley.

*Brehm, M., Imberman, S. A., & Lovenheim, M. F. (2015). *Achievement Effects of Individual Performance Incentives in a Teacher Merit Pay Tournament*. National Bureau of Economic Research.

*Briggs, D., Diaz-Bilello, E., Maul, A., Turner, M., & Bibilos, C. (2014). *Denver ProComp Evaluation Report: 2010-2012*.

Burgess, S., & Ratto, M. (2003). The role of incentives in the public sector: Issues and evidence. *Oxford Review of Economic Policy*, *19*(2), 285–300.

Chamberlin, R., Wragg, T., Haynes, G., & Wragg, C. (2002). Performance-related pay and the teaching profession: A review of the literature. *Research Papers in Education*, *17*(1), 31–49.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*(5), 1352–1370.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2011). Teacher Mobility, School Segregation, and Pay-Based Policies to Level the Playing Field. *Education, Finance, and Policy*, *6*(3), 399–438. doi:10.1162/EDFP_a_00040

*Contreras, D., & Rau, T. (2012). Tournament incentives for teachers: Evidence from a scaled-up intervention in Chile. *Economic Development and Cultural Change*, *61*(1), 219–246.

*Cooper, S. T., & Cohn, E. (1997). Estimation of a frontier production function for the South Carolina educational process. *Economics of Education Review*, *16*(3), 313–327.

*Cowan, J., & Goldhaber, D. (2015). Do bonuses affect teacher staffing and student achievement in high-poverty schools? Evidence from an Incentive for National Board Certified Teachers in Washington State. *Center for Education Data & Research*.

*Dee, T. S., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management*, *23*(3), 471–488.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of Human Resources*, *37*(4), 696–727.

Dowling, J., Murphy, S., & Wang, B. (2007). The effects of the career ladder program on student achievement. *Evaluation Report. Phoenix, AZ: Sheila Murphy Associates*.

Eggers, D., & Calegari, N. C. (2011, April 30). The High Cost of Low Teacher Salaries. *The New York Times*. Retrieved from http://www.nytimes.com/2011/05/01/opinion/01eggers.html

Evenden, E. S. (1919). *Teachers' salaries and salary schedules in the United States, 1918-19*. Washington, The National Education Association.

*Figlio, D. N., & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics*, *91*(5), 901–914.

*Fryer Jr, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). *Enhancing the efficacy of teacher incentives through loss aversion: A field experiment*. National Bureau of Economic Research.

*Fryer, R. G. (2011a). *Teacher incentives and student achievement: Evidence from New York City public schools*. National Bureau of Economic Research.

Fulbeck, E. S. (2014). Teacher Mobility and Financial Incentives: A Descriptive Analysis of Denver's ProComp. _Educational Evaluation and Policy Analysis_, _36_(1), 67–82.

*Glazerman, S., Protik, A., Teh, B., Bruch, J., Max, J., & others. (2013). _Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment_. Mathematica Policy Research.

*Glazerman, S., & Seifullah, A. (2012). An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Final Report. _Mathematica Policy Research, Inc._

*Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. _American Economic Journal: Applied Economics_, _2_(3), 205–227.

Goldhaber, D., DeArmond, M., Player, D., & Choi, H.-J. (2008). Why Do So Few Public School Districts Use Merit Pay? _Journal of Education Finance_, _33_(3), 262–289.

*Goldhaber, D. & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. _Economics of Education Review_, _31_(6), 1067–1083.

*Goodman, S., & Turner, L. (2011). Does Whole-School Performance Pay Improve Student Learning? Evidence from the New York City Schools. _Education Next_, _11_(2), 67–71.

Guthrie, J. W., Springer, M. G., Rolle, R. A., & Houck, E. A. (2007). Modern education finance and policy. Mahwah, NJ: Allyn & Bacon.

Hanushek, E. A. (2003). The Failure of Input‐based Schooling Policies. _Economic Journal_, _113_(485), F64–F98. doi:10.1111/1468-0297.00099

Harvey-Beavis, O. (2003). Performance-based rewards for teachers: A literature review. In *paper distributed at the third workshop of Participating Countries on OECD's Activity Attracting*. OECD: Athens, Greece.

Hatry, H. P., & Greiner, J. M. (1984). Issues in Teacher Incentive Plans. The Urban Institute. Washington D.C.

Heinrich, C. J., & Marschke, G. (2010). Incentives and Their Dynamics in Public Sector Performance Management Systems. *Journal of Policy Analysis and Management*, *29*(1), 183–208.

Hill, C.J., Bloom, H.S., Black, A.R., and Lipsey, M.W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, 2(3), 172-177.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

Hough, H. J. (2012). Salary Incentives and Teacher Quality: The Effect of a District-Level Salary Increase on Teacher Recruitment."

*Imberman, S. A., & Lovenheim, M. F. (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics*, *97*(2), 364–386.

Jinnai, Y., & others. (2016). The Effects of a Teacher Performance-Pay Program on Student Achievement: A Regression Discontinuity Approach. *Economics Bulletin*, *36*(2), 993–999.

Johnson, S. M., & Papay, J. P. (2010). Expecting Too Much of Performance Pay? *School Administrator*, *67*(3), 22–27.

Klassen, R. M., & Chiu, M. M. (2011). The occupational commitment and intention to quit of practicing and pre-service teachers: Influence of self-efficacy, job stress, and teaching context. *Contemporary Educational Psychology*, *36*(2), 114–129. doi:10.1016/j.cedpsych.2011.01.002

Ladd, H. F. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review*, *18*(1), 1–16.

LaFee, S. (2003). Professional Learning Communities. *School Administrator*, *60*(5), 6–12.

Lam, R. W., & Kennedy, S. H. (2005). Using metaanalysis to evaluate evidence: practical tips and traps. *The Canadian Journal of Psychiatry*, *50*(3), 167-174.

*Lavy, V. (2002). Evaluating the effect of teachers’ group performance incentives on pupil achievement. *Journal of Political Economy*, *110*(6), 1286–1317.

*Lavy, V. (2009). Performance pay and teachers’ effort, productivity, and grading ethics. *The American Economic Review*, *99*(5), 1979–2021.

Lazear, E. P. (1998). *Personnel economics for managers*. New York, NY: Wiley.

Lazear, E. P. (2001). Paying Teachers for Performance: incentives and selection. *Unpublished paper*, *Hoover Institution and Graduate School of Business*, *Stanford University*.

Lazear, E. P., & Shaw, K. L. (2007). Personnel economics: The economist’s view of human resources. *The Journal of Economic Perspectives*, *21*(4), 91–114.

Mark W. Lipsey, & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage publications.

*Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., & Epstein, S. (2011). *A big apple for educators: New York City’s experiment with schoolwide performance bonuses: Final evaluation report*. Rand Corporation.

*Martins, P. S. (2009). Individual teacher incentives, student achievement and grade inflation.

Mehta, J. (2013). *The Allure of Order: High Stakes, Dashed Expectations, and the Quest to Remake American Education*. New York, NY: Oxford University Press.

*Mizala, A., & Romaguera, P. (2005). Teachers' salary structure and incentives in Chile. In E. Vegas (Editor), *Incentives to Improve Teaching* (103-150). WorldBank: Washington D.C.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269.

*Muralidharan, K., & Sundararaman, V. (2009). *Teacher performance pay: Experimental evidence from India*. National Bureau of Economic Research.

Murnane, R., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, *56*(1), 1–18.

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, *26*(4), 909–949.

Podgursky, M. J., & Springer, M. G. (2010). Market-and performance-based reforms of teacher compensation: A review of recent practices, policies, and research. *Program on Education Policy and Governance at Harvard Kennedy School*.

Protsik, J. (1995). History of Teacher Pay and Incentive Reforms. Madison, WI: Consortium for Policy Research in Education. Retrieved from http://eric.ed.gov/?id=ED380894

*Proctor, D., Walter, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011). Making a Difference in Education Reform: ProComp External Evaluation Report 2006-2010. *Prepared for the Denver Public Schools*.

*Ritter, G., Holley, M., Jensen, N., Riffel, B., Winters, M., Barnett, J., & Greene, J. (2008). *Year Two Evaluation of the Achievement Challenge Pilot Project in the Little Rock Public School District*. Department of Education Reform, College of Education and Health Professions.

Ritter, G. and Barnett, J.H. (2013). A Straightforward Guide to Teacher Merit Pay: Encouraging and Rewarding Schoolwide Improvement. California: Corwin Press.

*Santibañez, L., Martinez, J. F., Datar, A., McEwan, P. J., Setodji, C. M., & Basurto-Davila, R. (2007). Breaking Ground: Analysis of the Assessment System and Impact of Mexico's Teacher Incentive Program. *RAND Corporation*.

*Schacter, J., & Thum, Y. M. (2005). TAPping into high quality teachers: Preliminary results from the Teacher Advancement Program comprehensive school reform. *School Effectiveness and School Improvement*, *16*(3), 327–353.

*Schacter, J., Thum, Y. M., Reifsneider, D., & Schiff, T. (2004). The Teacher Advancement Program Report Two: Year Three Results from Arizona and Year One Results from South Carolina TAP Schools. Santa Monica, CA: Milken Family Foundation. doi:10.1.1.419.7163

*Slotnik, W. J., Smith, M. D., Glass, R. J., Helms, B. J., & Ingwerson, D. W. (2004). Catalyst for Change: Pay for Performance in Denver Final Report. *Boston: Community Training and Assistance Center*.

*Slotnik, W., Smith, M., Helms, B., & Qiao, Z. (2013). It's More Than Money: Teacher Incentive Fund–Leadership for Educators' Advanced Performance, Charlotte-Mecklenburg Schools. *Boston, MA: Community Training and Assistance Center*.

*Sojourner, A. J., Mykerezi, E., & West, K. L. (2014). Teacher Pay Reform and Productivity Panel Data Evidence from Adoptions of Q-Comp in Minnesota. *Journal of Human Resources*, *49*(4), 945–981.

Springer, M.G. (2011). *Establishing a Framework for Evaluation and Teacher Incentives: Considerations for Mexico*. Paris: Organisation for Economic Co-Operation and Development.

Springer, M.G. and Balch, R. (2010). Design Components of Incentive Pay Programs in the Education Sector. In S. Sclafani (ed.), *Teacher Incentives and Stimuli*. Paris: Organisation for Economic Co-Operation and Development.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2015). Effective Teacher Retention Bonuses Evidence From Tennessee. *Educational Evaluation and Policy Analysis*, *38*(2), 199-221.

*Springer, M. G., Ballou, D., & Peng, A. X. (2014). Estimated Effect of the Teacher Advancement Program on Student Test Score Gains. *Education, Finance, and Policy*, *9*(2), 193–230.

*Springer, M. G., Pane, J. F., Le, V.-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S., & Stecher, B. (2012). Team Pay for Performance Experimental Evidence From the Round Rock Pilot Project on Team Incentives. *Educational Evaluation and Policy Analysis*, *34*(4), 367–390.

*Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., & Stecher, B. M. (2011). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness*.

*Springer, M. G., Lewis, J. L., Ehlert, M. W., Podgursky, M. J., Crader, G. D., Taylor, L. L., & Stuit, D. A. (2010). District Awards for Teacher Excellence (DATE) Program: Final Evaluation Report. *Policy Evaluation Report. Nashville, TN: National Center on Performance Incentives*.

*Springer, M. G., Podgursky, M. J., Springer, C., Hamilton, L. S., Lopez, O. S., Peng, A. X., & Stecher, B. M. (2009a). Texas Educator Excellence Grant (TEEG) Program.

*Springer, M. G., Lewis, J. L., Podgursky, M. J., Ehlert, M. W., Taylor, L. L., Lopez, O. S., & Peng, A. (2009b). Governor’s Educator Excellence Grant (GEEG) Program: Year Three Evaluation Report. *National Center on Performance Incentives*.

*Springer, M.G. and Taylor, L.L. (2016). Designing Incentives for Public School Teachers: Evidence from a Texas Incentive Pay Program. *Journal of Education Finance*, 41(3), 344-381.

Springer, M. G. (Ed.). (2010). *Performance incentives: Their growing impact on American K-12 education*. Brookings Institution Press.

Steele, J. L., Murnane, R. J., & Willett, J. B. (2010). Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. *Journal of Policy Analysis and Management*, *29*(3), 451–478.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology*, *53*(11), 1119-1129.

Stucker, J. P., & Hall, G. R. (1971). *The performance contracting concept in education*. DTIC Document.

Thoonen, E. E., Sleegers, P. J., Oort, F. J., Peetsma, T. T., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly*, *47*(3), 496–536.

Umansky, I. (2005). A literature review of teacher quality and incentives. *Incentives to Improve Teaching*, *21*. The World Bank.

Viscardi, D. (2014). The Teacher Pay for Performance Phenomenon. *Unpublished Doctoral Dissertation. Seton Hall University*.

*Wellington, A., Chiang, H., Heallgren, K., Speroni, C., Herrmann, M., Burkander, P., & others. (n.d.). *Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Three Years*. Mathematica Policy Research.

*Winters, M., Greene, J. P., Ritter, G., & Marsh, R. (2008). The Effect of Performance-Pay in Little Rock, Arkansas on Student Achievement. Working Paper 2008-02. *National Center on Performance Incentives*.

*Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review*, *30*(3), 404–418.

* denotes primary studies used in meta-analysis

Table 1

Quality Criteria for Assessing Risk of Bias

Critical Evaluation Criteria	***Quality Rating Considerations***
Was the intervention clearly defined?	Was the study a randomized control trial?
Were the research question(s) for this study clearly stated, and did the subsequent investigation answer the question(s)?	Was implementation fidelity measured and adequately described, and what are the implications of implementation fidelity on outcomes?
Did the study provide a clear review of prior research?	What are the relative strengths of the study design?
Were the inclusion and exclusion criteria for being in the study specified and applied uniformly to all participants?	Was the analytic approach adequately described, and what are the relative merits of the approach used?
Was a sample size justification or power description provided?	Was the comparison condition adequately described, and does the comparison group provide a reasonable counterfactual?
Did the study clearly define a control or comparison condition?	Were threats to internal and external validity considered and addressed?
Did the analytic strategy adjust statistically for confounding variables?	Were findings robust to different analytical decisions and model specifications?
Was the analytic strategy clearly defined and appropriate for answering the stated research questions?	Was baseline equivalence established between treatment and comparison groups? (This is unnecessary for some approaches such as the difference-in-difference design.)
Did the study include additional analyses (robustness/sensitivity checks) and subgroup analyses or adjust analyses?	What sampling decisions were made by the authors and did the analytic sample present any concerns to internal or external validity?
Were the outcome measures clearly defined, valid, reliable, and implemented consistently across study participants?	
Was the timeframe sufficient so that one could reasonably expect to see an association between exposure and outcome if it existed?	
Did the study contain specific objectives or hypotheses?	
Was the study population clearly specified and defined?	
Did the authors address study limitations, sources of bias, impressions and, if relevant, other limitation such as multiplicity of analysis?	
Was the interpretation of results consistent with estimates, and did the author consider other relevant evidence?	

Note: In the critical evaluation approach, studies meeting 13 out of 15 criteria were considered low risk of bias. In the quality rating approaching, studies with a rating of four or five out of five were considered low risk of bias.

Table 2

Descriptive Information on Primary Studies by Study and Program Characteristics

	Full Sample	**Randomized Control Trial**	**U.S. Only**
Study Characteristics			
Publication year	1997-2016	2004–2015	1997-2016
Peer reviewed	45% (18 studies)	50% (5 studies)	36% (10 studies)
Randomized control trial design	25% (10 studies)	100% (10 studies)	25% (7 studies)
Average Sample Size	594,751 students 9,254 schools	14,024 students 6,406 schools	737,474 students 5,641 schools
Range of Sample Size	323-8,561,194 students 92-43,251 schools	323-126,416 students 297-8442 schools	323-8,561,194 students 92–40,393 schools
Program Characteristics			
Range of treatment duration	1–12 years	1–4 years	1–12 years
Range of award receipt	$26 – $20,000	$169 – $15,000	$200 – $20,000
Rank-order tournament	25% (10 studies)	20% (2 studies)	13% (5 studies)
Group incentive structure	33% (13 studies)	60% (6 studies)	29% (8 studies)
Merit pay + other	43% (17 studies)	10% (1 studies)	57% (16 studies)
Merit pay + training	18% (7 studies)	10% (1 study)	21% (6 studies)
Public announcement of results	3% (1 study)	0% (0 studies)	0% (0 studies)
Award type	Gifts, one time bonuses, and salary increases	One time bonuses and salary increases	One time bonuses and salary increases
Number of Studies	40	10	28

Note: Merit pay + other refers to whether merit pay was implemented in conjunction with other reforms such as additional training. M*erit pay + training* refers to merit pay program that was implemented in conjunction with a training/professional development component. The full sample of 40 “studies” includes one that is an average of four reports or articles on the School-wide Performance Bonus Program and another that is an average of two reports on the Teacher Advancement Program. The percentages shown represent the percent of studies within each category.

Table 3

Meta-Analytic Results of the Effect of Merit Pay on Student Test Scores

Model	*Panel A: Main effect estimates*					*Panel B: Heterogeneity of study effects*			
	N	Effect Estimate	Standard Error	Lower Bound	Upper Bound	I^2	T^2	Cochrane Q	p_Q
Overall Effect									
All subjects	40	0.052	0.008	0.037	0.068	89.564	0.001	373.696	<.001
United States only	28	0.035	0.007	0.021	0.050	88.554	0.001	235.890	<.001
By Subject									
Math	33	0.066	0.009	0.048	0.084	92.830	0.002	446.304	<.001
ELA	27	0.037	0.007	0.022	0.051	88.782	0.001	231.779	<.001

Note: Assumed correlations between multiple, within-study outcomes is 0.5. Not all studies presented results separated by subject.

Table 4

Meta-Analytic Results of Moderators of the Effect of Merit Pay on Student Test Scores

Model	*Panel A: Main effect estimates*					*Panel B: Heterogeneity of study effects*			
								Cochrane	
	N	Effect Estimate	Standard Error	Lower Bound	Upper Bound	I^2	T^2	Q	P_Q
Overall Effect									
All subjects	40	0.052	0.008	0.037	0.068	89.564	0.001	373.696	<.001
By Study Characteristics									
Peer-reviewed	18	0.106	0.017	0.072	0.140	87.412	0.004	135.045	<.001
Randomized	12	0.074	0.019	0.036	0.112	76.131	0.003	46.084	<.001
By Program Characteristics									
Group incentive	13	0.111	0.027	0.058	0.164	94.776	0.007	229.704	<.001
Rank-order tournament	10	0.063	0.027	0.010	0.115	91.438	0.005	105.118	.019
Merit pay+other	17	0.044	0.009	0.026	0.061	77.416	0.001	70.847	<.001
Merit pay+training	7	0.041	0.023	-.004	0.086	73.968	0.002	23.048	.076

Note: Assumed correlations between multiple, within-study outcomes is 0.5. *Merit pay + other* refers to whether merit pay was implemented in conjunction with other reforms such as additional training. *Merit pay + training* refers to merit pay program that was implemented in conjunction with training/professional development component.

Table 5

Risk of Bias and Unit of Analysis

Model	*Panel A: Main effect estimates*					*Panel B: Heterogeneity of study effects*			
	N	Effect Estimate	Standard Error	Lower Bound	Upper Bound	I^2	T^2	Cochrane Q	p_Q
Main Effect									
All	40	0.052	0.008	0.037	0.068	89.564	0.001	373.696	<.001
Risk of Bias									
Critical evaluation approach	36	0.043	0.007	0.029	0.058	89.121	0.001	321.716	<.001
Quality rating approach	18	0.057	0.015	0.027	0.086	86.584	0.003	126.717	<.001
Unit of Analysis									
School level	9	0.066	0.024	0.019	0.114	86.630	0.004	59.835	<.001
Student level	29	0.051	0.010	0.031	0.072	90.067	0.002	281.899	<.001

Note: Assumed correlations between multiple, within-study outcomes is 0.5. Not all studies presented results separated by subject.

Table 6

Summary of Literature on Effect of Merit Pay on Teacher Recruitment and Retention

Program Evaluated	Study Design Study Period Sample	Intervention	Results
Mostly Positive Effects			
North Carolina Bonus Program (Clotfelter et al., 2008)	Hazard Models 2001-02 through 2003-04 29,562 teachers	Teachers were awarded up to $1,800 in schools serving low-income or low-performing students.	The bonus award reduced turnover rates by about 17% among targeted teachers.
Missouri Career Ladder Program (Booker & Glazerman, 2009)	Cox Proportion Hazard Models, IV 1989-90 to 2006-07 (Program started in 1986) Over 140,000 teachers	Teachers meeting state and district-level performance criteria could receive supplementary bonus pay ($1,500-$3,000) for completing academic responsibilities.	Teachers in participating districts were less likely (.85 times) than teachers in non-participating districts to move to a different district. Teachers in participating districts were also less likely (90 percent as likely) to leave teaching.
Texas District Awards for Teacher Excellence Program (DATE) (Springer et al., 2010)	Probit and Multinomial Logit Models, District Fixed Effects 2002-03 to 2008-09 (Program began in 2008-09.) Over one million teacher-year-school observations	Eligible school districts that decided to accept the grants (ranging from $4,395 to over $13,000,000 in 2006-07) created incentive pay plans either for all schools or for select, high-need schools.	Turnover rates were 1.3 percentage points lower for schools in districts with district-wide DATE incentive plans. Individual teacher awards greater than $100 were associated with a significant decrease in the probability of teacher turnover under district-wide DATE plans.
Talent Transfer Initiation (TTI) (Glazerman et al., 2013)	Block (teacher-team) randomization, OLS (2009-10 through fall 2011-12) 10 districts 114 schools	Highest-performing teachers (top 20 percent) were offered $20,000 to transfer to low-performing schools for 2 years.	During the years of implementation, retention rates for TTI teachers were significantly higher than non-TTI teachers. Retention rates were not significantly different after the intervention ended.
Denver Professional Compensation System (ProComp) (Fulbeck, 2014)	Hazard Models 2001-01 through 2010-11 (Program began in 2005-06) 12,952 teachers	Teachers were eligible for a ten different financial incentives ranging from $376 to $3,379.	Receipt of a ProComp incentive was associated with a 30% decrease in teachers' odds departure from a school.
Washington Challenging Schools	RD 2007-08 to 2012-13	National Board Certified Teachers were awarded up to $5,000 at high	The CSB bonus increased the proportion of National Board Certified teachers by about

Bonus (CSB) (Cowan & Goldhaber, 2015)	(Program started in 1999-00.) Over 200,000 student-year observations	poverty schools.	0.022, and the bonus increased the probability that a newly hired teachers is National Board Certified by 1 percentage point.
Some Positive Effects			
Texas Educator Excellence Grant (TEEG) Program (Springer et al., 2008)	Probit and Multinomial logit models 2002-03 to 2007-08 (Program implemented in 2006-07 to 2009-10.) Over one million teacher-year-school observations	Schools with a high percentage of economically disadvantaged students were awarded one-year grants ranging from $40,000 to $295,000 per year. Schools were required to dedicate 75% of these funds to classroom teacher performance awards.	Schools participating in the TEEG program did not experience any systematic reduction in teacher turnover during 2007 and 2008, but the probability of turnover fell as the size of the individual teachers' award increased such that awards of $3,000 per teacher reduced the predicted turnover rate among recipients to less than a third of the predicted turnover rate observed before the TEEG program.
Texas Governor's Educator Excellence Grant (GEEG) (Springer et al., 2009)	Probit and Multinomial logit models, campus fixed effects 2002-03 to 2007-08 (Program occurred in 2005-06 to 2007-08.) Over one million teacher-year-school observations 99 schools	Schools with a high percentage of economically disadvantaged students were awarded three-year grants ranging from $60,000 to $100,000 per year. Schools were required to dedicate 75% of these funds to classroom teacher performance awards.	Teacher turnover was consistently lower (3.21 percentage points) in GEEG schools than non-GEEG schools in the first year, but this difference did not persist. The actual receipt and size of the award had a strong impact on teacher turnover. Experienced teachers receiving an award of $1,250 or more had a significantly lower probability of turnover.
California Governor's Teaching Fellowship (GTF) (Steele et al., 2010)	IV, Hazard Models 1998-99 though 2002-03 (Program occurred in 2000-01 through 2001-02) 27,106 teacher candidates	Teachers were awarded $20,000 for four years of service in a low-performing school.	The award increased recipients' probability of teaching in a low-performing school by about 28 percentage points, but there was no difference in the probability of school exits between recipients and non-recipients.
Quality Teacher and Education Act in San Francisco Unified School District (QTEA) (Hough, 2012)	DID 2002-03 through 2010-11 (Program began in 2008-09) 1,363-8,939 teacher-year observations.	Teachers received salary increases of $500-6,300 and retention bonus ranging from $2,500-3,000 and a $2000 bonus for working in hard-to-staff schools.	The QTEA retention bonus had no significant effect on rates of teachers staying the same school or transferring, but in hard-to-staff schools, retention rates were 15 percentage points higher for teachers targeted by the retention bonus than in the absence of QTEA.
Chicago Teacher Advancement Program	School Randomization, Propensity Score Matching	Teachers were eligible for extra pay and increased responsibilities based on	Chicago's TAP had a positive, significant impacts on year-to-year retention for the first

(TAP) (Glazerman & Seifullah, 2012)	(Program occurred in 2007-07 to 2009-10.) 781-2,694 teachers	their contribution to student achievement and classroom observations.	two years of the program, but not for the third year. Retention rates for TAP schools ranged from zero to 20 percentage points higher than matched comparison schools, depending on cohort and year.
Minnesota Quality Compensation (Q Comp) (Choi, 2015)	Fixed Effects Model 2002-03 through 2009-10 (Program began in 2005-06) 12,708 observations in 1,734 schools	One component of the Q Comp plan required participants to include a performance-base pay component tied to student achievement and teacher evaluations.	Q Comp implementation had no significant association with changes in teacher retention overall, but schools implementing Q Comp for five years have teacher retention rates that were 6.32 percentage points higher than schools without five years of Q Comp implementation.
Tennessee Governor's Retention Bonus Program (Springer et al., 2015)	RD 2011-12 through 2013-14 (Program occurred in 2013-14) 56 schools, 587 teachers	Highly effective teachers received a $5,000 retention bonus for working in low-performing priority schools.	No significant effects on teacher retention overall. However, tested-subject teachers who received the bonus were 20% more likely to remain in a priority school than teachers who did not receive a bonus.
Statistically Insignificant Effects			
School-wide Performance Bonus Program (Fryer, 2013)	School Randomization IV (2008-2010) 40-187 schools 4,693-21,700 teachers	Schools meeting performance standards received $3,000 per union represented staff member	No significant effect for teacher retention at the school or district level.
IMPACT Teacher Evaluation System in Washington D.C. (Dee & Wyckoff, 2015)	RD 2009-10 through 2011-12 2,132-4,178 teachers	Teachers meeting certain performance ratings were eligible for bonuses up to $25,000 and base pay salary increases of up to $27,000.	A highly-effective rating and eligibility for financial incentives raised teacher retention by 3 percentage points, but this result is statistically insignificant.

Notes: OLS stands for Ordinary Least Squares. IV stands for Instrumental Variables. DID stands for difference-in-difference, RD stands for regression discontinuity.

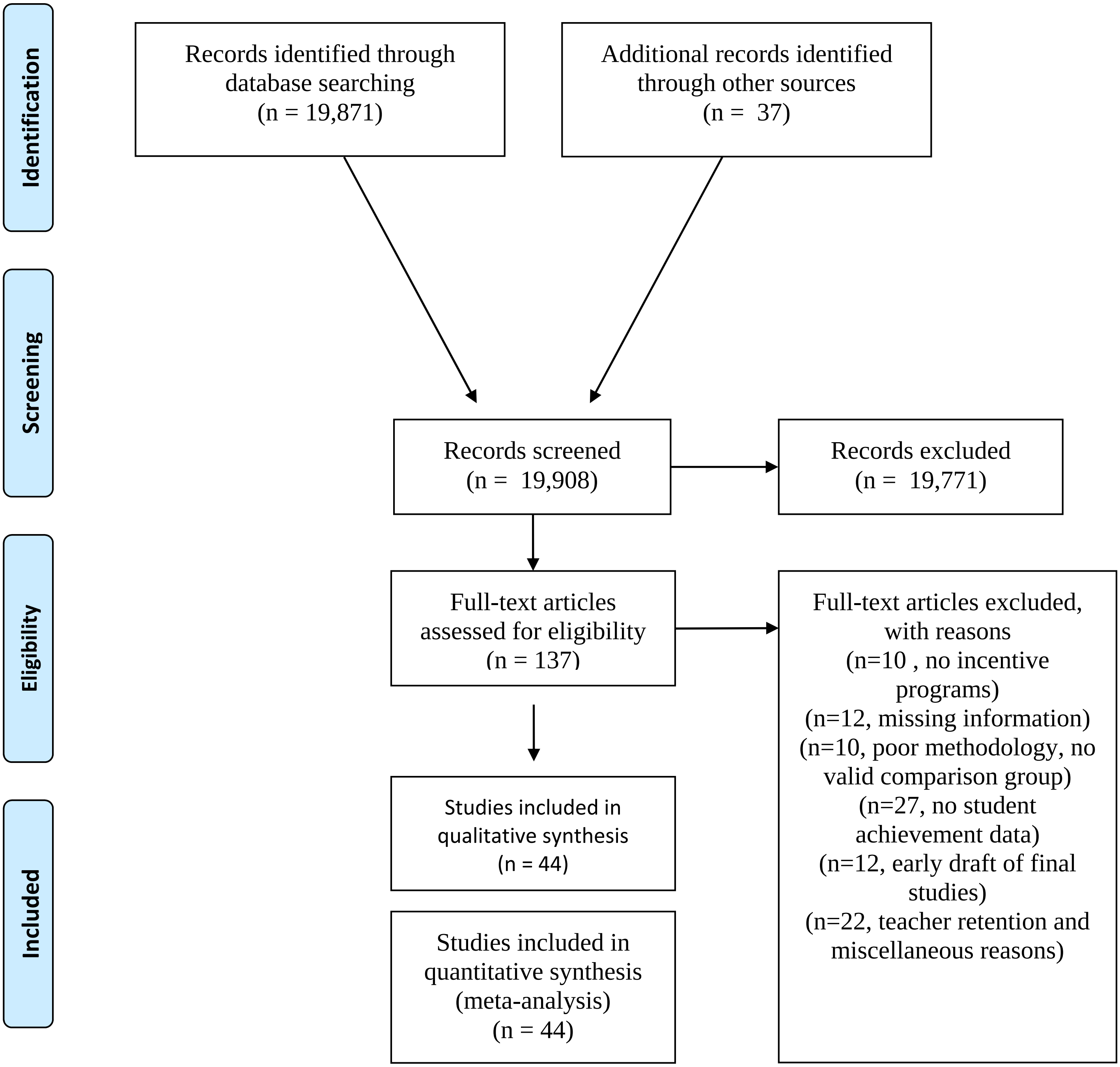
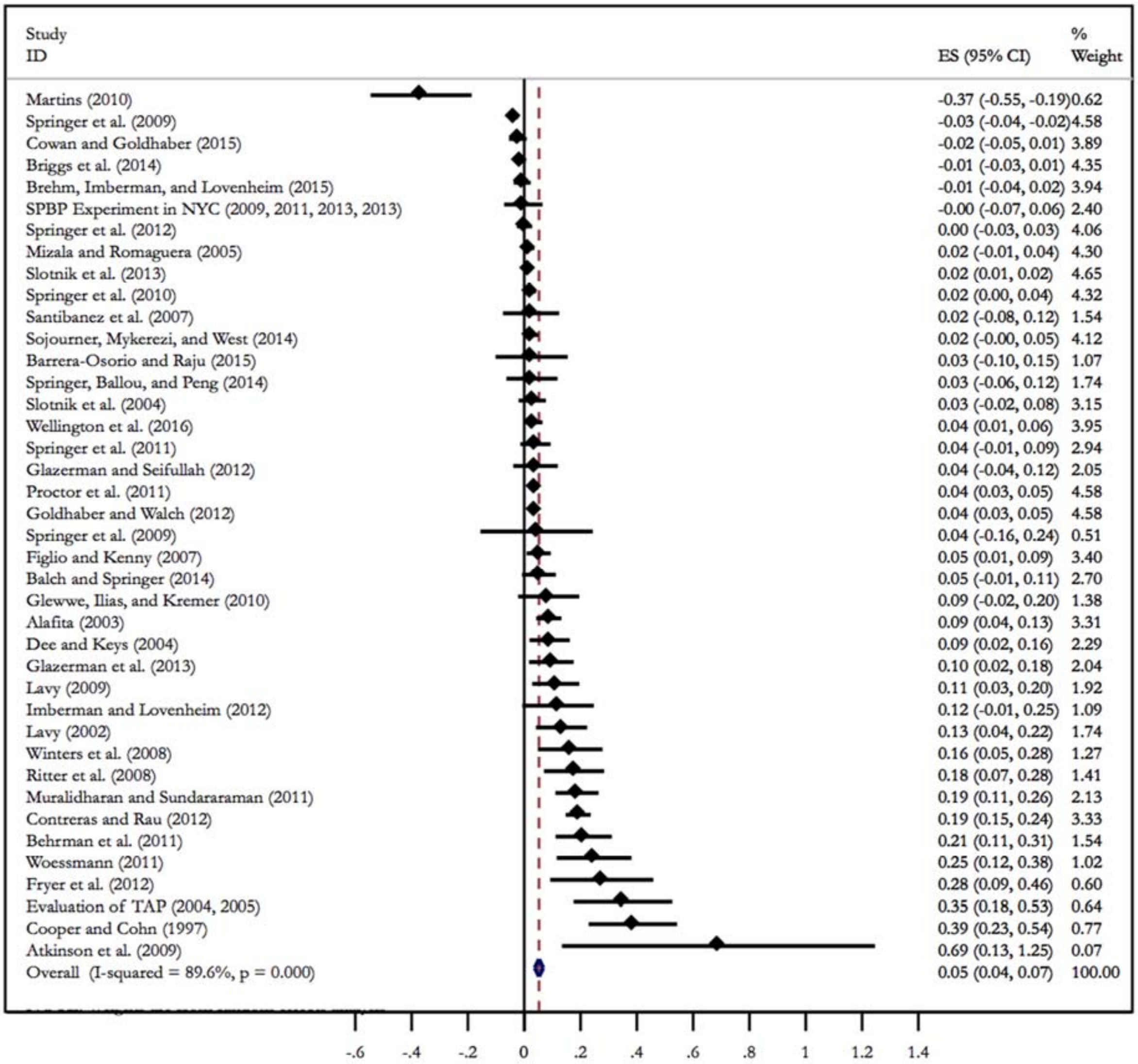

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 19,871)
Additional records identified through other sources (n = 37)
Records screened (n = 19,908)
Records excluded (n = 19,771)
Full-text articles assessed for eligibility (n = 137)
Full-text articles excluded, with reasons
(n=10 , no incentive programs)
(n=12, missing information)
(n=10, poor methodology, no valid comparison group)
(n=27, no student achievement data)
(n=12, early draft of final studies)
(n=22, teacher retention and miscellaneous reasons)
Studies included in qualitative synthesis (n = 44)
Studies included in quantitative synthesis (meta-analysis) (n = 44)

Figure 1. Flow diagram depicting the literature screening process resulting in the final sample of primary studies included in the quantitative analysis. Adapted from Moher et al. (2009).



Note: Weights are from random effects analysis. ID is identification. ES is effect size. CI is confidence interval. Correlation between multiple outcomes within a study, r, is 0.5

Figure 2. Forest plot for overall effect estimates of merit pay programs on student test scores from primary studies.

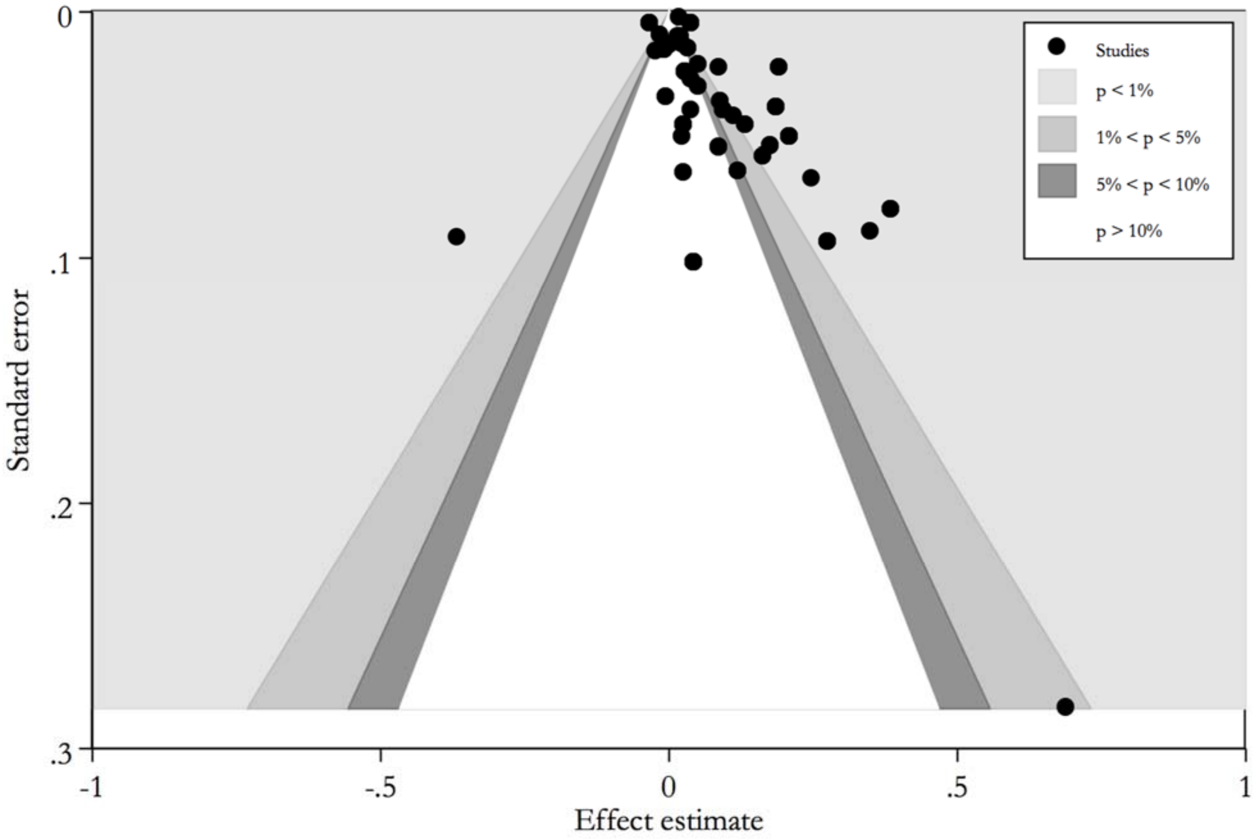


Figure 3. Contour enhanced funnel plot with one effect estimate per primary study

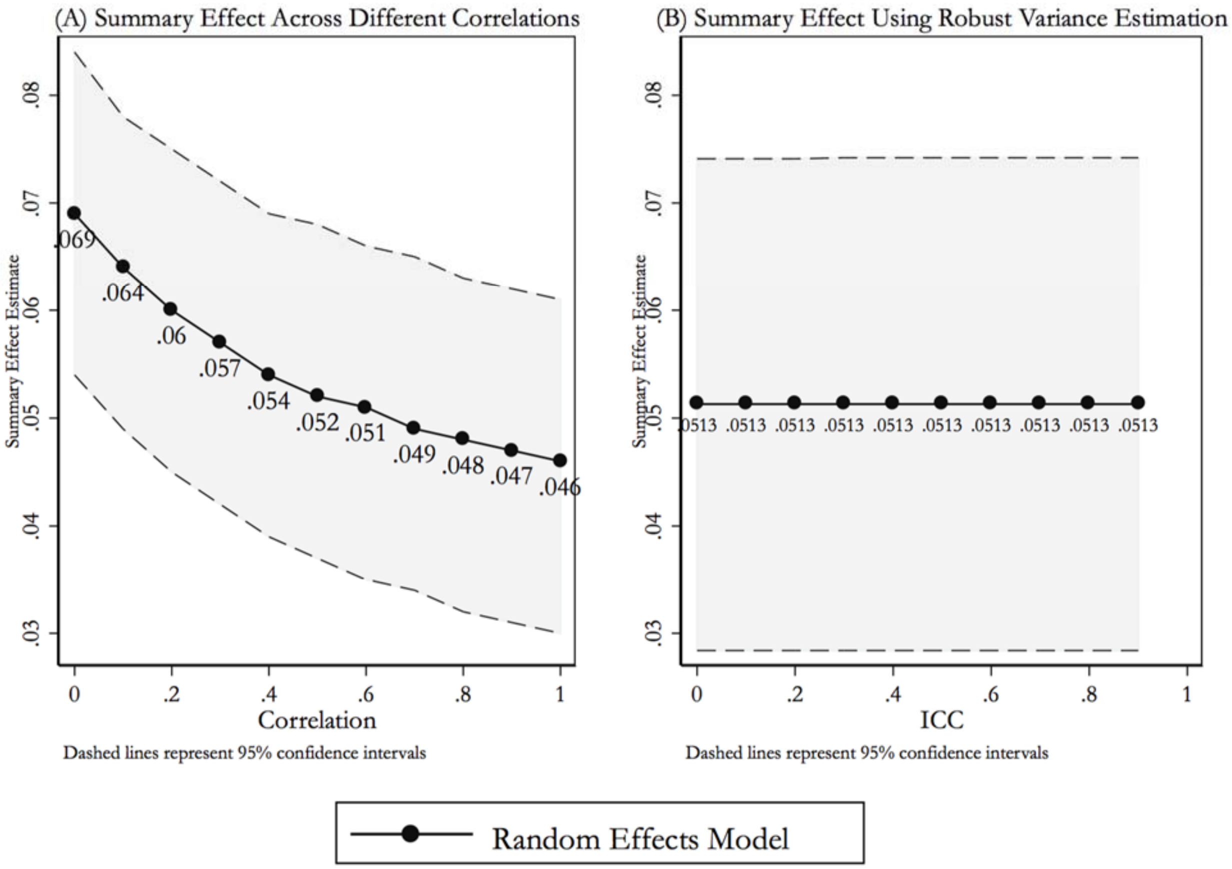


Figure 4. Sensitivity of overall effect estimate to (A) Different correlations between multiple within-study outcomes (r_{ij}) and (B) Different correlations (ρ) using robust variance estimation

Supplemental Tables

Table S1

Results by Database

Database	Results
Directory of Open Access Journal (DOAJ)	5761
WorldCat (incudes dissertation and theses)	3869
Taylor and Francis Online	1463
ProQuest	1378
Wiley Online Library	1102
Google Scholar	1000
JSTOR	1000
SpringerLink	987
ERIC	758
SciVerse Science Direct	681
NBER	357
Web of Science	327
Project MUSE	309
EconLit	208
ProQuest Dissertations and Theses	206
Education Full Text	145
OneFile (GALE)	120
PsycINFO	94
SAGE	72
Sociological abstracts	34
Total	19871

Table S2

Coding Guide

Study Characteristics

Variable	**Description**	**Operationalization**
id	ID Number assigned to study	Continuous
leadauth	Name of lead author	Nominal
title	Title of paper	Nominal
yearpub	Year paper was published	Continuous
pubtype	Type of publication (academic journal, policy report, conference paper, etc.)	Nominal
rct	Randomized control trial indicator	0, 1
randomized	Did this study intend to randomize whether students/teachers/schools were placed into control vs. experiment groups?	0,1
evaluate	Is this study an evaluation of an existing intervention?	0,1
peer review	Is the study a peer-reviewed publication?	0,1
working paper	Is it a working paper?	0,1
use	Is the study based in the U.S.?	0,1
otherctry	Name of the country where the study was conducted if not the U.S.	Nominal
state	Name of state (if U.S.=1)	Nominal
depvar	The dependent variable(s) of the study: test score, teacher value-added scores, teacher attendance, etc.	Nominal
design	Type of design: pre-post randomized control group, pre-post nonrandomized, post-test only matched samples, etc.	Nominal
confdsgn	Coder’s confidence in validity of the design	Uncertain, somewhat certain, certain (0, 1, 2)
program	Name of program/experiment	Nominal
lngthrt	Length of merit pay program in years	Continuous
lvlrandm	Level of randomization (district, school, teacher, etc.)	Nominal
eqvscores	Pre-test score equivalence between treatment and control	0,1
eqvgrade	Grade equivalence between treatment and control	0,1
eqvfrpl	Free and reduced price lunch (FRPL) equivalence between treatment and control	0,1
eqvsubject	Academic subject equivalence	0,1
eqvrace	Race equivalence between treatment and control	0,1
noneqv	List any characteristics that were statistically significantly different between treatment and control	Nominal
confidenceeqv	Coder’s confidence in equivalence of the treatment and control groups	Uncertain, somewhat certain, certain (0, 1, 2)

component	Indicator of whether merit pay was a component of some larger program, such as a district wide initiative that also included teacher mentors	0,1
district	Name of district	Nominal
comparison	Comparison group (business as usual, BAU, or other intervention)	1-BAU, 0-Other
gradelvl	Grade level of the intervention	Categorical
minamnt	Minimum amount of merit pay received	Continuous
maxamnt	Maximum amount of merit pay received	Continuous
avgamnt	Average amount paid per teacher	Continuous
paytype	Pay type: bonus, change in salary bracket, gift	Nominal
recurring	Can a teacher receive the bonus more than once?	0,1
competition	Was merit pay a competition such that some teachers receiving a bonus meant others could not?	0,1
recipient	Did the teachers directly receive the pay?	0,1
paycriteria	Criteria teachers to receive merit pay: test scores, observation ratings, multiple evaluative measures, etc.	Nominal
publiconly	Public school only indicator (1-public only, 0-private only or both public/private)	0,1
training_received	Indicator for whether teachers received any training as part of the merit pay program	0,1
public_announce	When teachers received a merit pay incentive was there a public announcement for the bonus winners?	0,1
groupincentive	Did the teachers receive incentives as part of a group?	0,1

Study Outcomes

Variable	**Description**	**Operationalization**
year	Year(s) in which study was conducted	Continuous
outcometype	Test score or pass rate	Nominal
math	Outcome is math test scores	0,1
reading	Outcome is reading/ELA test scores	0,1
schlevel	School level: elem, ms, hs, k-8	Categorical
lvlmeasure	Level of measure: student, school, or teacher	Nominal
instrument	Name of test administered to students (e.g., Iowa Test of Basic Skills)	Nominal
sig10	Outcome is significant at 10% level	0,1
sig5	Outcome is significant at 5% level	0,1
sig1	Outcome is significant at 1% level	0,1
method	Analytical design (e.g., OLS, RD, Diff in Diff, IV)	Nominal
standardized	Were the coefficients standardized?	0,1
sd	Standard deviation of the dependent variable	Continuous
beta	Regression coefficient from regressing outcome on merit pay	Continuous
stdbeta	Standardized beta coefficient	Continuous
avgbeta	Average if multiple standardized beta coefficients are reported from in one study	

se	Standard error of beta coefficient	Continuous
stdse	Standard error of standardized beta coefficient	Continuous
stdvar	Standardized variance (*stdse* squared)	Continuous
tstat	*t* statistic if reported	Continuous
r_sq	R^2 if reported	Continuous
samplesize	Sample size	Continuous
avgsamplesize	Average sample size if multiple outcomes are reported	Continuous
prevscores	Lagged dependent score if included in regression	0,1
frpl	Binary indicator if FRPL was included as covariate	0,1
sped	Binary indicator if SPED was included as covariate	0,1
lep	Binary indicator if LEP was included as covariate	0,1
pct_frpl	Binary indicator if % FRPL was included as covariate	0,1
pct_sped	Binary indicator if % SPED was included as covariate	0,1
pct_lep	Binary indicator if % LEP was included as covariate	0,1
notes	Additional notes about the study	Qualitative notes