THE CAMPBELL COLLABORATION

Merit Pay Programs for Improving Teacher Retention, Teacher Satisfaction, and Student Achievement in Primary and Secondary Education: A Systematic Review

Joshua Barnett, Gary Ritter, Nathan Jensen, Wen-Juo Lo, & George Denny

Submitted to the Coordinating Group of:

☐	Crime and Justice
☒	Education
	☐ Disability
☐	International Development
	☐ Nutrition
☐	Social Welfare
☐	Other:

Plans to co-register:

☒	No		
☐	Yes	☐ Cochrane	☐ Other
☐	Maybe		

Date Submitted:
Date Revision Submitted:
Approval Date: May 4, 2014
Publication Date: July 1, 2014

Note: Campbell Collaboration Systematic Review Protocol Template version date: 24 February 2013

BACKGROUND FOR THE REVIEW

Approximately 95 percent of K-12 teachers in the U.S. work in a school or district with a salary schedule that provides salary increases largely based on years of experience and number of degrees attained (Podgursky & Springer, 2007). This “single-salary schedule” or “lock-step schedule” was developed to address equity issues and has spread across the globe, where teachers most often receive increased compensation based on qualifications and years of service (inputs) rather than on measures of actual teaching and their students’ performance (outputs) (Glazerman, Chiang, Wellington, Constantine, & Player, 2011).

Historically, elementary teachers were often paid less than secondary teachers, and women and minority teachers earned lower salaries than white, male teachers. Thus, the single-salary schedule was developed and implemented to pay the same salary to teachers with the same qualifications (such as post-secondary degrees earned) and years of experience, regardless of a teacher’s race, gender, or grade level taught (Odden & Kelley, 2002). In this way, the justification for paying differential salary amounts was objective, measurable, and not subject to administrative whim.

The development of the single-salary schedule was intended to foster fairness and equity in pay. Ironically, many critics of the single-salary schedule claim that this compensation approach creates inequities for teachers, as highly-effective teachers are paid the same as less effective teachers in such a system (in whatever way “effectiveness” is defined) (Ravitch, 2010). Further, some reformers argue that the single-salary approach results in inequities for students as well, especially poor or minority students, since schools cannot offer greater salaries to higher-quality teachers to attract them to work with these types of students (Hanushek, 2013).

Critics of the single-salary system have also argued that current practices related to teacher pay and employment have little, if any, consequences associated with poor performance (Cohn & Teel, 1991; Dee & Keys, 2004; Lazear, 2002). This issue is not only due to the single-salary schedule, but also because of the tenure rules employed in most school districts across the nation. For instance, in other fields where employees do not benefit from the same level of job security afforded to teachers through the tenure process, poor job performance can lead to the loss of employment (Lazear, 1996). However, this is generally not the case in the field of teaching; indeed, even today, when dire budget circumstances in some states are leading to teacher layoffs, the layoff decisions are generally based on teacher seniority rather than on some measure of teacher performance (Goldhaber, 2011; National Council on Teacher Quality, 2010).

Perhaps most fundamentally, in the single-salary system, compensation is seldom based on any evidence that a teacher is effective at enhancing student learning (Ballou & Podgursky, 1997); this system assumes that teaching ability improves with more years of experience and higher degrees. However, many researchers have found that additional degrees do not result in enhanced student learning, and that the benefits of teacher experience plateau after several years (e.g., Goldhaber, 2002; Hanushek, 2007). If this finding is true, the current teacher compensation structure is not likely to move the field forward in terms of raising student achievement or ensuring that each child has an effective teacher, as teachers will get raises each year according to the salary schedule, regardless of whether or not their work with students actually merits an increase in pay.

Further, in the current single-salary system, effective teachers have little opportunity to have their work recognized or rewarded, and may seek out additional “compensation” through alternative means. For example, teachers may seek out placements in high-income schools with student populations they view as easier to educate, or they may leave the classroom altogether and enter school administration as a way to earn additional pay. Or, teachers may simply leave the field of education, and pursue work in a field in which they receive greater financial rewards commensurate with their job performance. In each of these instances, the outcome is the opposite of what decision-makers would want – the single-salary structure incentivizes effective teachers to move further away from the students who likely need them the most (Belfield & Heywood, 2007; Figlio & Kenny, 2006; Ritter & Barnett, 2013).

Because of this, an increasing number of schools have begun to adjust how their teachers are compensated, and are now providing teachers with the opportunity to earn financial bonuses for demonstrating exceptional work with their students.

Commonly referred to as merit-pay, this compensation approach provides teachers with additional financial compensation based on, among other things, how well their students perform on measures of student learning (outputs), rather than the traditional approach which compensates teachers based on years of experience or degrees attained (inputs). Ultimately, there are three general ways that a shift from the current single-salary system to a merit-based system of compensating teachers might impact teachers and students (Ritter & Barnett, 2013). First, by moving towards this type of system, teachers are provided with extra incentives (in the form of bonuses) to focus on raising student achievement, which could result in students showing greater progress over the course of the year. Second, having their hard work with their students recognized and rewarded may provide teachers with extra incentives to stay in that school and/or in the teaching profession, thereby leading to greater retention of high-quality teachers. Finally, because this type of compensation approach is focused on identifying and rewarding effectiveness, such a system may begin to draw talented new individuals into the teaching profession who are confident they could be successful, know their hard work will be recognized, and are more comfortable being held accountable for their work (since they are entering into a system with a high level of accountability already in place).

Advocates of merit pay believe that teachers who prove themselves to be effective should be given opportunities to earn appropriate rewards while remaining in the field of education and in the classroom where they can directly impact student learning (Figlio & Kenny, 2006;

Glewwe, Ilias, & Kremer, 2003; Ritter & Barnett, 2013). Furthermore, proponents of equity might argue that the most effective teachers should be encouraged, not discouraged, to work with students in low-income areas with the greatest educational needs. Thus, in this age of increasing accountability for teachers and schools, compensating teachers based on what they actually do with their students, in lieu of the more traditional approach, has become increasingly common in education today.

However, there is much debate about whether merit pay programs actually do lead to positive benefits for teachers and students, or if they instead lead to a number of negative outcomes for teachers (Cordes, 1983; Sawchuck, 2010). For example, one oft-cited concern is that merit pay programs can have a negative impact on teachers, students, and the overall culture of a school. Kohn (1993) argued that merit pay bonuses can be viewed as punishments instead of rewards, since teachers will constantly be worried about being caught for doing something wrong, as opposed to being praised for doing something right. Further, Kohn suggested that if not every teacher receives a bonus, then relationships between teachers could potentially be damaged, and may result in teachers feeling the need to compete with their peers instead of support them (Kohn, 1993).

The idea that teachers would stop collaborating with each other, and in turn begin to compete for a finite pool of bonus money, is one of the primary sources of opposition towards this type of compensation strategy (Goldhaber et al., 2008). This opposition can perhaps be best summarized by the following statement from the National Education Association's (NEA), the largest teachers association in the United States:

> "Merit pay systems force teachers to compete, rather than cooperate. They create a disincentive for teachers to share information and teaching techniques. This is especially true because there is always a limited pool of money for merit pay. Thus, the number-one way teachers learn their craft—learning from their colleagues—is effectively shut down. If you think we have a turnover problem now, wait until new teachers have no one to turn to."

There is also concern that merit pay is ill-suited for use in schools, as it is difficult to truly measure the impact teachers have on student learning (Goldhaber et al., 2008; Murnane & Cohen, 1986; Ramirez, 2011; Sawchuck, 2010). This is because there are myriad factors that can influence student performance on standardized assessments on a given day (such as the child being sick, or missing breakfast that morning, etc.) (Berliner, 2010; Papay, 2011). If a teacher's "effectiveness" is measured based on these measures of student achievement, then the actual impact the teacher had on his or her students may not be accurately captured (Darling-Hammond, 1986, 2006). Further, teachers may be unfairly categorized as ineffective based on student test results that do not accurately reflect the actual work they have done with their students. Thus, despite the potential positive impacts of merit pay, there is also the potential for a number of negative outcomes as well.

Regardless, in light of problems with recruiting high quality teachers into the classroom, retaining them in the profession, and holding them accountable for student achievement, more and more states and school districts (e.g., Charlotte-Mecklenburg, Cincinnati, Denver, Douglas County, Nevada, Los Angeles, Texas, and Washoe County) and federal governments are moving towards the use of merit pay, in hopes of recruiting and retaining more qualified teachers who can improve student performance (Kelley, 1998, 2000; Odden & Kelley, 1997). Additionally, the discourse related to teacher recruitment and retention has made its way into the education preparation programs across the globe, where much attention is now being given to how teachers need to be trained (Cochran-Smith, Feiman-Nemser, McIntyre, & Demers, 2008; Darling-Hammond, 1986, 2006; Shulman, 1988) and prepared to be evaluated once in the field (Eckert, 2009).

As the conversation about merit pay and teacher evaluation continues to unfold, U.S. President Obama has endorsed the investigation of salary reforms, including the use of merit pay, through the Race to the Top competition (awarding approximately $4.5 billion) and the Teacher Incentive Fund competition (approximately $1 billion). Yet, as national policymakers and school leaders consider the use of merit pay, they often find themselves confronted with a number of core complications inherent in the creation of such plans. As discussed in the extant evidence, three key methodological issues make compensation reform challenging:

- First, what evaluation instruments should be used to determine which teachers should receive greater compensation? Most merit pay advocates insist that teacher “merit” should be based, at least in part, on standardized measures of achievement or growth for students in a teacher's classroom. Not surprisingly, the idea of holding teachers accountable for student test score gains is a source of great debate and discord (Amrein-Beardsley, 2012; Harris, 2011). Nevertheless, most merit pay plans do and will include standardized measures of student test performance. However, school leaders and policymakers need to know more about which measures of student achievement and growth have been used and which ones are better suited for this type of compensation approach.

- Second, should rewards be based on the efforts of individual teachers or groups of teachers? Some merit pay plans today are school-based; that is, the school receives some sort of rating for its overall performance over a given time period (Springer & Winters, 2009). Then, all the teachers in that school receive a bonus based on the “merit” of the school as a whole. Other plans rate individual teachers based on their individual classroom performance and allocate different reward levels to different teachers (Barnett, Ritter, Winters, & Greene, 2007). However, school leaders do not have consistent evidence on the impact of these different approaches.

- Third, how should different award levels or bonus amounts be determined for different school personnel? In schools or districts with the most limited merit pay programs,

> only those teachers who teach “core” subjects with corresponding standardized assessments can participate. In other plans, non-core teachers (such as art and music teachers) and school support staff (such as custodians and aides) are eligible for awards, although the award levels and rating systems might be different. Additionally, some merit pay plans include awards for school administrators based on overall performance of the school or district. Therefore, school leaders need information on which types of educators and school employees to include in the bonus pool, and what levels of rewards to offer to each type of employee.

In general, the challenge for compensation reform is that the theory makes intuitive sense, but the methodological and practical issues associated with implementing such a system are complicated in that no single merit pay “plan” exists; instead, there are numerous ways in which school leaders have implemented a merit pay plan for their teachers. The details of these plans are important for understanding and determining the overall impact of merit pay as a compensation strategy.

Yet, current research on the effectiveness of merit pay for promoting positive impacts for teachers and students is unfortunately mixed; consequently, research evidence has historically contributed little to the discussion on the merits of merit pay or to identifying important components from which practitioners might build a program. Several foundational studies and more recent panels of studies have also contributed to the confusion over the policy of merit pay. In particular, the historical review by Moore-Johnson (1984) contended that merit-based reforms had been put forward in the 1920s, 1950s, and 1980s with little evidence showing changes to teachers’ motivation levels. More broadly, Murnane and Cohen (1986) expressed concerns that historical applications of market-based reforms have produced negligible positive outcomes in the educational sphere.

More recently, the Economic Policy Institute gathered a who’s who list of education researchers to examine how teachers are evaluated (Baker et al., 2007); the report concluded that “there are good reasons for concern about the current system of teacher evaluation” (p. 1) and discusses various teacher evaluation metrics, surmising that “any sound evaluation will necessarily involve a balancing of many factors” (p. 1). In response to this work, the Brookings Brown Center gathered an equally distinguished group of education researchers (Glazerman et al., 2010) who discussed the need for student achievement calculations to be a part of a teacher’s evaluation and broader compensation conversations.

Harvey-Beavis (2003) reviewed the evidence in international education and determined that the theory of merit pay is strong, but the application in practice has not been realized. This review specifically examines the available academic and policy literature from English-speaking countries, with the majority of research coming from programs in the United States. The paper provides an overview of different types of reward programs, arguments for merit pay programs, arguments against merit pay programs, reasons why merit pay programs are difficult to implement, and a summary of the current evidence. Chamberlin et al. (2002) conducted a review of evidence in British schools and reached similar conclusions to those of Harvey-Beavis, noting that a number of programs were implemented but the evaluations of such programs were limited or poorly performed. Their paper specifically notes the challenges to implementation and the difficulties in isolating effects.

Podgursky and Springer (2007) conducted a review of the evidence on merit pay and noted that results from evaluations of these programs were mixed but showed promise for impacting student achievement. Their review concludes by noting that even though "the empirical literature is not sufficiently robust to prescribe how systems should be designed – for example, optimal size of bonuses, mix of individual versus group incentives – it does make a persuasive case for further experiments by districts and states, combined with rigorous, independent evaluations" (p. 910). In the recent review of evidence on merit pay, Yuan et al. (2013) examined three randomized studies and concluded that teacher motivation was not affected by having the opportunity to earn a merit pay bonus.

The value and need for a more rigorous systematic review is that over the previous five years merit pay programs have expanded within education policy circles, and more and more districts, states, and nations are trying different approaches to compensation reform. Specifically, the infusion of resources from the U.S. Department of Education through the Teacher Incentive Funds have allocated over $500 million across over 100 locations to develop, implement, and evaluate teacher compensation reforms. The majority of these resources were allocated in the TIF Cycle 3 and Cycle 4 grants, which were provided in 2009 and 2012, respectively. As such, a thorough review of existing research is needed, with particular attention given how these programs impact teachers and students, which evaluation tools are useful in measuring teacher "merit", how group or individual rewards impact teacher response to these programs, and what is known about programs aimed towards different types of teachers and personnel.

OBJECTIVES OF THE REVIEW

To begin understanding the objectives of this review, we provide the following definition for how we are examining "merit pay". Consistent with prior literature (Podgursky & Springer, 2007), we define merit pay as rewards for individual teachers, groups of teachers, or schools on any number of factors, including student performance, classroom observations, and teacher portfolios. Merit-based pay is a reward system that hinges on student outcomes attributed to a particular teacher or group of teachers rather than on inputs such as skills or years of experience. These rewards can be bonus amounts provided at the end of the year, or they can be increases to base salary.

The general goals of merit pay programs are fairly straightforward: a) incentivize teachers to invest greater time and effort into their teaching, b) reward teachers for exceptional work, and c) encourage teachers to stay in the classroom. However, school districts that implement such a program are often left developing them with little knowledge of the effectiveness or characteristics of other systems. The objectives of the proposed review are to answer the following questions:

1. To what degree do merit pay programs impact student achievement outcomes?
2. To what degree do merit pay programs impact teacher outcomes (e.g. retention; satisfaction)?
3. What are the distinguishing characteristics of the most successful merit pay programs?
4. For which subgroups of students/teachers/ school systems are merit pay programs most or least beneficial?

The objectives of this review are intended to inform the practical issues school leaders and policymakers face when considering a merit pay program and provide guidance on the impact of these programs on student achievement.

METHODS

In the following methods section, we describe operationally how we will conduct this review.

Criteria for Inclusion and Exclusion of Studies in the Review

Types of Studies

We will only consider studies for this review that have been published since 2000, as the systematic review of Podgursky and Springer (2007) and Harvey-Beavis (2003) only located two studies prior to 2000 (Clotfelter & Ladd, 1996; Ladd, 1999). Further, federal policies have dramatically shaped the scope of merit pay programs since 2000, with the passage of No Child Left Behind (2002) and the reauthorization of the Elementary and Secondary Education Act (2007). Federal grant programs such as the Teacher Incentive Funds, Teacher Quality Partnerships, and Race to the Top have also provided unparalleled resources in recent years for districts and states to consider compensation reform.

The study must be written in English, but the study may be conducted in the United States (including the 50 states, the District of Columbia, territories, and tribal entities) or other nations. We are aware of merit pay programs operating in Australia, New Zealand, India, Israel, and other nations and will incorporate these studies into our review if they meet inclusion criteria. For example, Muralidharan and Sundararaman (2009) assessed the impact of a merit pay program in India, Lavy (2002; 2009) conducted evaluations of merit pay programs in Israel, and Glewwe, Ilias, and Kremer (2003) evaluated a large-scale merit pay program in Kenya. This international research can greatly contribute to the understanding of how merit pay programs affect teachers and students, and will be included in our review; however, due to the limitations of the reviewers, the research must be available in English to be included.

The literature review is limited to only include reviews of merit pay programs implemented in public schools (including public charter schools) instead of private schools, and only research focused on merit pay programs implemented in K-12 education settings. Due to the organization, policy, and requirement differences in education environments between private schools and pre-school or post-secondary schools as compared to public K-12 schools, research focused on merit pay programs in private schools will be excluded from this review.

One important consideration for this review is that all research should be focused on merit pay programs in which teachers/school personnel earned year-end bonuses based on some annual evaluation of their performance, including (but not limited to) measures of how individual teachers specifically impacted student achievement, school-wide achievement gains for employees who do not directly impact student achievement (such as school counselors), or end-of-year evaluations by a principal, etc. The rationale for this follows the prior work of Podgursky and Springer (2007), in which the authors stated that, "Merit-based pay rewards individual teachers, groups of teachers, or schools on any number of factors, including student performance, classroom observations, and teacher portfolios" (p. 912).

Evaluations of performance could vary by program for this review, as could the exact structure of the individual merit pay programs. For example, some merit pay programs provide rewards only to teachers of core subjects, whereas other programs financially reward all employees in a school. In some programs, bonuses are fixed (where a teacher either earns a bonus of $3,000, or nothing at all) or continuous (where a teacher could earn any amount up to $3,000), and bonuses can either be distributed to individual personnel based on individual performance, or the same bonus can be given to all teachers based on the performance of the school as a whole. Programs might also be structured in a zero-sum fashion, such that only a limited number of teachers earn a bonus (such as the top ten teachers in a school), whereas other programs are specifically designed to ensure that everyone in a school benefits financially under the program. In all of these examples, the exact structure of the program differs, but at their core these types of programs are all designed to provide financial rewards to school personnel based on some measure of their performance each year, and all would still be included in this review.

By contrast, these criteria also mean that programs under which personnel can earn additional money for non-performance based activities, such as programs in which teachers earn an annual bonus simply for completing additional professional development activities, would not be included in this literature review. For example, this review does not include any evaluations of career ladders programs despite the fact that these programs are often associated with merit pay. Briefly, in career ladder programs such as those used currently in Arizona and previously in Missouri schools, teachers are/were able to receive supplementary pay for meeting certain performance criteria. These criteria can include extra teaching work or participating in professional development activities, and often also include meeting some form of tenure requirement. Thus, while teachers are rewarded under these programs for meeting performance criteria, the criteria used in these programs are not inherently focused on capturing the impact these teachers have on their students on an annual basis. Put differently, teachers in career ladder programs are doing extra work for extra money, but their extra work is not evaluated for quality or impact on student achievement.

Finally, one of the primary goals of this review is to identify high-quality research specifically aimed at evaluating the impact of merit pay. Because of this, one of the key criteria in this review process is to only include research that features an evaluation component, where the exact impact of the use of merit pay on some outcome measure (such as student achievement, teacher attitudes, etc.) could be directly quantified or measured relative to a comparable alternative standard or counterfactual. This guideline is to ensure that the research used for this review includes actual evaluations of merit pay, rather than opinions for or against the use of this compensation strategy, or simply discussions about various aspects of the use of merit pay.

Study Design

We will include two types of study designs in this review: randomized control trials and quasi-experimental trials. The review will not include pretest/posttest, single-subject, or qualitative-only studies. For a study to meet design specifications it must include a treatment group (receiving merit pay) compared to a no treatment group (not receiving merit pay); that is, we will only include studies where a control group has an absence of any intervention (e.g. business-as-usual). Quasi-experimental studies that employ treatment and control groups matched on pretests of key outcome variables (where baseline equivalence can be determined) will be included in this review.

We will incorporate the What Works Clearinghouse (WWC) Standards 3.0 (Institute of Education Sciences, n.d.) regarding study design designations. Specifically, in a randomized control trial, study participants (students, teachers, classrooms, or schools) must have been placed into “each study condition through random assignment or a process that was functionally random (such as alternating by date of birth or the last digit of an identification code). Any movement or nonrandom placement of students, teachers, classrooms, or schools after random assignment [is performed] jeopardizes the random assignment design of the study” (p. 12).

In a quasi-experimental design, the intervention group includes participants who were either self-selected (for example, volunteers for the intervention program) or were selected through another process, along with a comparison group of nonparticipants. Because the groups may differ, a quasi-experimental design must demonstrate that the intervention and comparison groups are equivalent on observable characteristics.

Following the standard established by the WWC Handbook 3.0 (n.d.), baseline equivalence must be demonstrated on observed characteristics related to student achievement. Specifically, the difference between the treatment and comparison groups must be less than 0.25 standard deviations to be retained and determined "equivalent" (p. 14). If the difference between treatment and comparison groups is between 0.05 and 0.25 standard deviation units, then a statistical adjustment must be made (e.g., OLS regression adjustment; fixed effects; ANCOVA) consistent with the procedures defined by the WWC (which are drawn from Ho, Imai, King, & Stuart, 2007).

With regard to attrition, we will examine overall and differential attrition. Following the WWC Standards 3.0 (n.d.), which uses a "pessimistic but still reasonable" assumption of bias from attrition, we will retain studies with overall attrition rates below 60% and differential attrition below 12% (p. 12).[1]

Types of Participants and Interventions

Eligible interventions to be included are merit pay programs "where merit pay is defined as rewards for individual teachers, groups of teachers, or schools on any number of factors, including student performance, classroom observations, and teacher portfolios. Merit-based pay is a reward system that hinges on student outcomes attributed to a particular teacher or group of teachers rather than on "inputs" such as skills or knowledge" (Podgurksy & Springer, 2007, p. 912). Again, these bonuses can be provided as either end-of-year bonuses or as increases to base salary.

Participants for this review include K-12 teachers and their corresponding students. A merit pay program's effectiveness could vary by subgroups of students, teachers, or schools. However, whether a study examines effects on subgroups does not affect the inclusion of the study for review.

The merit pay program must also meet the following criteria:

- The intervention must be carried out in a K-12 (or the equivalent in other countries) school.
- The intervention must include K-12 teachers. Programs aimed towards staff development of principals/administrators will be included as long as they also include a focus on teachers.
- The intervention must include differential pay based on a defined performance measure.

[1] The WWC Handbook 3.0 does not specify an upper limit for "high attrition" but provides over 6.3 as an upper limit when overall attrition is low. The sliding scale found on Table III.1 (page 12) of Handbook 3.0 and Figure A1 on page 34 of Handbook 2.1 provide guidance on the trade-offs between overall and differential attrition. For our review, we determined that any study with greater than 12% differential or 60% overall would contribute to bias in the results of the study.

- The intervention must be implemented for at least one academic- or school-year (may be implemented in schools with quarter, semester, or year-round calendars).
- The intervention must include a specified financial reward (e.g. additional pay; bonus) that is calculated and provided for each school year.
- The comparison groups (of students and teachers) will receive no treatment or business-as-usual (e.g. traditional salary as indicated in their district or state guidelines).
- Career ladder programs where teachers advance on a set performance ladder, will not be included in the study, as the programs more closely approximate the traditional single-salary system than a merit or bonus system.

Types of Outcome Measures

The study needs to include at least one measure that involves direct assessment of student achievement or growth or a teacher outcome (such as retention or measures of job satisfaction). Additional student outcomes (e.g., attitude towards school, motivation, and self-efficacy) are not the focus of this review and do not qualify as relevant outcome measures. Given the nature of most district and state testing standards in the United States under the No Child Left Behind Act of 2001, the authors anticipate student achievement outcomes to be based on achievement in math and reading; however, other tests (e.g., locally developed exams) will be included provided they demonstrate sufficient reliability and validity.

Reliability will be assessed using the following standards specified by WWC Version 3.0 standards: internal consistency (minimum of 0.50), temporal stability/test-retest reliability (minimum of 0.40), and inter-rater reliability (minimum of 0.50) (p. 15). Over-alignment issues will also be considered, as outcome measures should not be too closely linked to the program; an example of this would be student scores improving on an assessment that is connected to a merit-pay reward system, while student scores from other assessments not connected to the merit-pay system remain unchanged. Consistent with the WWC standards, if data are not available to evaluate the reliability and validity of a measure, and we cannot determine if the measure has face-validity, we will exclude the outcome. For teacher outcomes, retention or satisfaction/attitudes may be examined; however, teaching measures must also demonstrate sufficient reliability and validity as noted under the WWC standards. The reviewers will calculate effect sizes for each measure in an attempt to make the data as comparable as possible.

In addition to the results must either include an effect size or include enough information for us to be able to calculate an effect size. In cases where the reported data seem inconsistent with other information reported, we will exclude the study. For example, if an ANOVA was conducted on two groups with a total sample of 55 participants, but the reported degrees of freedom were 1,92, then we would exclude this study from our review.

Search Strategy and Rationale for Identification of Relevant Studies

To provide context for how school personnel might respond to merit pay programs, and to assess what types of achievement gains might (or might not) be expected as a result of such programs, we seek to identify research that addresses the impact merit pay programs have on students, teachers, and school personnel. To ensure that the review of existing research is as comprehensive as possible, we will first develop criteria to help focus our search of merit pay research.

For these purposes then, the guidelines used to identify merit pay research will adhere to the following search criteria:

- Research conducted within the previous twelve years (since January 1, 2000);
- Focused on merit pay programs implemented in public schools (including public charter schools), not private schools;
- Focused on merit pay in K-12 education;
- Must be focused on merit pay programs in which teachers/school personnel earn year-end bonuses based on some evaluation of their performance;
- The research includes an evaluation component specifically aimed at measuring the impact of merit pay on teachers, school personnel, and/or students.

Application of Selection Criteria

After developing the search criteria, the next step in the review is to apply these criteria to a number of different search options to identify as much high-quality merit pay research as possible. For the purposes of this review, we will use the following search engines and alternative search options.

The primary means by which research will be identified is through searches of electronic databases, specifically Australian Education Index; British Education Index; CBCA Education; EBSCO Academic Search; Education Fulltext; EconLit; ERIC; Francis; ProQuest Dissertations & Theses; ProQuest Research Library; PsycInfo; and Web of Science.

In these databases, the following search terms will be used in combination to maximize the identification of relevant merit pay literature (including dissertations, working papers, reports, and journal articles):

> “merit pay” OR “performance pay” OR “teacher salar*” OR “teacher compensation” OR “salary scale” OR “teacher incentive*” OR “teacher bonus*” OR “pay-for-performance”

AND

evaluat* OR effective* OR outcome OR measure* OR quantif* OR “student achievement” OR “student performance” OR success*

AND

K12 OR “K-12” OR kindergarten OR “Grade 1” OR “Grade 2” OR “Grade 3” OR “Grade 4” OR “Grade 5” OR “Grade 6” OR “Grade 7” OR “Grade 8” OR “Grade 9” OR “Grade 10” OR “Grade 11” OR “Grade 12” OR “High School” OR “Elementary School” OR “Primary School” OR “Public School”

The search terms with asterisks (“effective*” and “evaluat*”) are included to capture all studies that use any variation of these terms – effective* will locate studies with the terms effectively, effectiveness, etc. In addition to the explicit search terms cited above, the researchers will consult each electronic database Thesaurus to locate additional terms. The researchers will also consult a university reference librarian to help confirm the search and Thesaurus usage was exhaustive and conducted appropriately.

To ensure that relevant articles on merit pay are not overlooked in our initial searches of the aforementioned databases, we will also conduct title reviews of every journal article since January 1, 2000 from six prominent education and economics journals, specifically the *Journal of Policy Analysis and Management, Education Finance and Policy, Educational Evaluation and Policy Analysis, Research in the Schools, Review of Education Research,* and *Journal of Public Economics*. Additionally, hand searches will be conducted of articles since January 1, 2000 from various education policy research organizations and think-tanks, such as the National Bureau of Economic Research, the National Center on Performance Incentives, the Rand Corporation, Mathematica Policy Research, and MDRC. Further, we will review the conference proceedings since January 1, 2000 of the Association for Education Finance and Policy and the National Education Finance Conference both of which have historically addressed issues related to teacher salary. These organizations were all identified based on discussions with researchers with significant experience in the field of merit pay. The purpose of these searches is to identify research on merit pay that has not been published in an academic journal, and thus might not be located in the previous search processes.

Specifically with regard to the gray literature search, researchers will locate the title list of each journal and organization publication. Any journal title or organization publication that uses one of the search strategy terms (e.g., merit pay; teacher pay; etc.) or is viewed to substantively address merit pay by the researcher will be retained.

We will also include all of the articles used by Harvey-Beavis (2003), Chamberlin et al. (2002), and (Podgursky and Springer (2007) in their reviews of merit pay (when these articles meet our review criteria). Here again, the goal is to ensure that all relevant research on the topic of merit pay is located and included. We will also consult with other key researchers in the field for studies related to the topic of which they are aware (e.g., Dr. Michael Podgursky, Dr. Matthew Springer, Dr. Roland Fryer, Dr. David Figlio, Dr. James Guthrie, Dr. Julie Marsh). As a final step in the search process, we will consult with the university reference librarian to help conduct a Web search using the Advanced Search form in Google and Bing to locate additional manuscripts. Similar to the other gray literature searching, we will retain any manuscripts using the search strategy terms or those that seem to substantively relate to the topic of merit pay.

Once the full list of titles is obtained through the search criteria, the authors will begin selecting the studies based on the inclusion/exclusion criteria. For this purpose, two reviewers will read each title and abstract for all collected evaluations/studies; this initial step is intended to remove those studies that clearly do not meet the eligibility criteria. During this initial read of titles and abstracts, if either reviewer considers a study to be eligible, then the study will be retained for further review. In the next pre-coding phase, two reviewers will read each full article, with extra attention given to the methods section, to determine if each article still meets all eligibility criteria. After these initial reviews, all articles that meet our eligibility criteria will be full reviewed and coded by two authors. Potentially eligible studies will be obtained from the University of Arkansas Library, Interlibrary Loan, ERIC, and online when full text is available.

Description of Methods Used in Primary Research

Typical evaluations of merit pay programs employ either experimental or quasi-experimental designs and include testing of treatment and control groups both before and after intervention. The most common outcome investigated is student achievement in reading and math, usually as measured by state- and/or district-level standardized test scores.

One study that exemplifies the methods commonly used in merit pay evaluations examined the city-wide merit pay program implemented in New York City. In the New York merit pay program, schools were randomly assigned to participate. If the students in a participating school achieved at a predetermined performance level, then the school received a lump sum equal to $3,000 per school employee to be distributed at the discretion of a committee of teachers and school personnel (i.e., evenly among all employees or differentially among various employees). This evaluation conducted by Fryer (2011) showed lower achievement for the students in merit pay program schools. Fryer noted that this program resulted in no instances of positive increases in student achievement and, at the middle school level, participating schools actually experienced lower student achievement than did the comparison schools.

Criteria for Determination of Independent Findings

Many studies report results along multiple outcome measures (i.e. language scores, math scores, attendance, teacher retention) and often include several variations of the same measurement (i.e. teacher, parent, and/or student reporting of achievement). Effect sizes for each measure will be extracted and coded into our analysis. The methods for maintaining statistical independence during analysis in cases where multiple effect sizes are available are described below.

Details of Study Coding Strategies

When studies provide it, we will collect information on student characteristics, including baseline achievement score, grade, gender, socioeconomic status, racial/ethnic composition, second-language status, and “at-risk” status (as defined by study authors). Reviewers will also collect information on teacher demographics (gender, ethnicity, etc.), tenure in teaching, and educational attainment when individual information is provided. School characteristics of interest for this review include location of participating schools, school type (public, private, religious), school SES (e.g., Title I school), average class size (small, medium, large), school size (small, medium, large), and school community (rural, suburban, urban). The quality of each study (and its reporting) will be assessed according to several characteristics, including: 1) the transparency of the study; that is, the clarity with which the investigators reported the assignment procedures; 2) the integrity of the assignment design and whether investigators address violations of the design; 3) the existence of high levels of attrition (particularly, differential attrition between treatment and control groups) from baseline samples to analysis samples; and 4) baseline equivalence for quasi-experimental designs and for experimental designs with high levels of attrition. A copy of the draft coding manual is included as Appendix A.

At least two reviewers will independently extract and code data from each full article in a coding guide that will include the following:

- Study citation & author affiliations
- Study sponsor & relation to program
- Peer-reviewed or non-peer-reviewed
- Program objective & rationale
- Program location & setting (i.e., urban vs. rural school)
- Time frame of program & study
- Sampling & assignment procedure (i.e., experimental, quasi-experimental)
- Student characteristics in both experimental and control groups (grade level, gender, ethnicity, & socioeconomic status)
- Program characteristics (i.e., evaluation instruments; eligible recipients; amount of bonus; group or individual components)

- Attrition of sample
- Analytical techniques (i.e., multiple regression analysis)
- Outcome measures used as indicators of student achievement (subgroups, if applicable)
- Results of the study
- Methodological weaknesses and criticisms of study designs
- Inclusion decision for systematic review (met criteria or not)

We will assess inter-rater agreement, or coding reliability, for all studies, and resolution of coding disagreements will be resolved by meeting and discussing contested items. Only data with perfect agreement will be entered for each study.

Statistical Procedures and Conventions

We expect that the analytic strategies discussed below will allow us to address the four research questions proposed within this review. Specifically, the treatment of effect sizes and multiple outcomes will allow us to isolate the impact of merit pay programs on student achievement outcomes. Through synthesizing the results of studies meeting our retention criteria, we will be able to also inform the conversation on how merit pay programs impact teacher outcomes, including retention and satisfaction. Further, with our coding strategies and the explanation of the separate outcomes and program characteristics accounted for throughout, we will be able to distinguish between the most relevant and highest yield components of various merit pay program approaches. As we have expressed throughout this review, merit pay programs are not individual systems; rather each program or plan created and evaluated is dependent upon a variety of decisions (Ritter & Barnett, 2013). Finally, the analytic strategy and coding structure put forward in this protocol will allow for subgroup analyses to be examined.

Effect Sizes and Analysis Model

Once we have identified and compared all outcomes measured in the included studies, we will select the appropriate effect size metric for the meta-analysis. We anticipate using the standardized mean difference effect size statistic (Cohen's *d*) with Hedges' correction (Hedges' *g*) to control bias due to small sample size for each outcome measure (Lipsey & Wilson, 2001). Ninety-five percent (95%) confidence intervals (CIs) will also be calculated for each effect size to examine whether the effect is statistically significant. When means and standard deviations are not reported, we will attempt to estimate the effect sizes using the procedures described by Lipsey and Wilson (2001). Analysis of impacts on retention will be reported both in percentage point differences as well as log-odds ratios, since these are typically dichotomous outcomes. In instances where these outcomes are reported in multi-level categorical levels, we will convert them into binary measures (retained/not retained).

The decision rules for the binary system will be developed upon an investigation of these data and explained in the final review.

Inverse variance weights will be employed when estimating average effects. Given the wide array of samples and methods expected to be included in any meta-analysis, we believe the random-effects model will be more appropriate than the fixed-effect model (e.g., Field, 2001, 2003; Hedges & Vevea, 1998; Hunter & Schmidt, 2000, 2004; National Research Council, 1992; Raudenbush, 1994). The fixed effect model assumes that variance among studies is due only to sampling error and variability can be presented as within-study variance, σ^2. In contrast, the random effect model is based on the assumption that the variability is due to both within-study sampling error and the variation across studies (i.e., between-study variance, τ^2). In other words, variation between studies could be systematic instead of assuming that all variability is due to sampling error. To assess the heterogeneity in the summary effect sizes, we will compute 1) the τ^2 estimate of between-study variance, 2) Cochran's *Q* statistics (Cochran, 1954), and 3) the I^2 statistics (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003). The *Q* test follows a chi-square distribution with n -1 degrees of freedom (n: the number of studies). If *Q* test is significant, it indicates that there are considerable differences existing between studies in the analysis and leads to a conclusion that there is heterogeneity among studies. However, one of well-known limitations by the *Q* test in the meta-analysis is that it overly detects very small variability when the number of studies is large and it also detects true heterogeneity poorly when the number of studies is small (Higgins, Thompson, Deeks, & Altman, 2003). Therefore, the I^2 index, (*Q*-[n-1]/*Q*), can overcome this shortcoming by calculating the ratio of true heterogeneity to total variation and determining the magnitude of heterogeneity (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Multiple Outcomes for Single Studies (and Heterogeneous Outcome Measures)

Based on prior reviews, the evaluation studies of merit pay programs employ a variety of different outcome measures to assess program effectiveness. Some evaluations use district or local language and/or math scores, other evaluations use state-level standardized assessments. Examples of standardized exams used include the Stanford Achievement Test (SAT-10) and the Iowa Test of Basic Skills (ITBS). Because the different studies employ different outcomes in different ways, it may not be prudent to calculate the "effect" of each individual study or the "overall effect" of all available studies. To determine whether a given intervention has a greater effect in any one area, we will conduct separate meta-analyses of key outcome areas, such as standardized overall reading results and standardized overall math results. If a study measures a key outcome in several ways, we will ensure that each study only contributes one data point to the analysis for each key outcome in order to ensure that no individual study is unduly "weighted" in the meta-analysis.

Unit of Analysis Issues

We expect that most studies will have the teacher or the classroom as the unit of analysis. For studies with the school as the unit of analysis, if there are at least five studies analyzed at the school level, we will do separate analyses for those studies. We won't be able to compare teacher-level and school-level studies due to the following reasons. First, Hedges (2007) indicated the information of intraclass correlation (ICC) needed for the computation of effect size estimates in the cluster-randomized design. However, we cannot locate any study or national report that provides the reasonable values of ICC in merit pay programs. Second, although we might conduct systematic search (e.g., Gulliford, Ukoumunne, &Chinn, 1999; Murry & Blitstein, 2003; Murry, Varnell, & Blitstein, 2004; Verma & Lee, 1996) to estimate the possible ICC value, we expect the criteria that we setup for our study may not fit in this type of searching. In other words, we need to conduct a separate study to be able to obtain the possible ICC value before we combine teacher-level and school-level studies, which is out of the scope of our study.

Publication Bias

In addition, we adopt the Duval and Tweedie (2000) funnel plot with the trim and fill analysis to investigate publication bias.

Moderator Analyses

When sufficient numbers of studies are found, the reviewers will also conduct subgroup analysis that compares the different results of subgroups of studies, including: 1) studies reporting on direct vs. indirect outcome measures; 2) studies of group vs. individual awards; and 3) studies of programs focusing on language vs. math or other academic subjects. As previously noted, we will then calculate effect sizes to attempt to compare the data between studies as best as possible.

For the categorical moderator variables, we will conduct several univariate moderator analyses. Initially, based on the moderator variables, we will split the overall studies (i.e., k) into p subsets (i.e., p repentant the number of subsets) and conduct separate meta-analysis on each subset. Then, we calculate the mean effect size for each subset across studies and the Q_W test (i.e., a within-group homogeneity statistic). Afterwards, we examine the between-group goodness-of-fit statistic (Q_b) with an approximate chi-square distribution with p-1 degree of freedom. If the result of Q_b test is significant, it indicates a significant moderator is presented. For the continuous moderator variables, we will use DerSimonian and Laird method (i.e., regression based on the method of moments weights) to test the potential continuous moderators. This procedure will be carried out by using SPSS macros (Lipsey & Wilson, 2001; Wilson, 2006).

Sensitivity Analysis

Using the Comprehensive Meta Analysis ® 2.2 software, we will test the extent to which our main results are sensitive to any one study's inclusion in the meta-analysis. The "one study removed” analysis presents the average standardized mean difference of all remaining studies after each study, in turn, is removed from the analysis. All meta-analysis will be conducted by first using the listwise deletion data with group comparisons on outcomes with missing effect sizes imputed, and second using the multiple imputation dataset with group comparisons on outcomes with missing effect size imputed. If the results are similar, the results based on the imputed dataset will be reported and the listwise deleted results will be reported in the Appendix because the former will have, depending on how severe the missing data, more statistical power. If the results are dissimilar, the results will be reported vice versa.

Missing Data

In the case of critical data not being reported in the studies, we will attempt to contact study authors for the information. If we are unable to find the critical information, we will retain them throughout the review process, but we will exclude such studies from our analysis.

Software and Resources

All study coding and data management will be done using Microsoft Excel. To conduct the meta-analysis, the authors will use Comprehensive Meta-Analysis software developed by Biostat. Other researchers with expertise in meta-analysis will be consulted throughout this process, as needed.

Treatment of Qualitative Research

This review focuses on studies involving randomized controlled trials or quasi-experimental designs; therefore, no qualitative studies will be coded for the purposes of this project.

REFERENCES

Amrein-Beardsley, A. (2012). Recruiting expert teachers into high-needs schools: Leadership, money, and colleagues. *Education Policy Analysis Archives, 20*(27). Retrieved from http://epaa.asu.edu/ojs/article/view/941

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. *Economic Policy Institute Briefing Paper #278*. Retrieved from http://www.epi.org/page/-/pdf/bp278.pdf

Ballou, D., & Podgursky, M. (1997). *Teacher pay and teacher quality*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Barnett, J. H., Ritter, G. W., Winters, M. A., & Green, J. P. (2007). Evaluation of year one the Achievement Challenge Pilot Project in the Little Rock School District. *Department of Education Reform Working Paper Series*. Retrieved from http://www.uark.edu/ua/der/der_research.php?sort=author#barnett

Berliner, D. C. (2010, June 29). New analysis of achievement gap: ½ x ½ = 1½. Washington Post. Retrieved from http://voices.washingtonpost.com/answer-sheet/guest-bloggers/new-analysis-of-achievement-ga.html#more

Belfield, C., & Heywood, J. (2007). Performance pay for teachers: Determinants and consequences. *Economics of Education Review*, 27, 243-252.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. NewYork, NY: Wiley.

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., & Ladd, H. (1996). Recognizing and rewarding success in public schools. In H. Ladd (Ed.), *Holding schools accountable: Performance-related reform in education* (pp. 23-63). Washington, DC: The Brookings Institution.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101-129.

Cochran-Smith, M., Feiman-Nemser, S., McIntyre, J., Demers, K. E. (2008). *Handbook of research on teacher education: Enduring questions in changing contexts.* (3rd ed). New York: Routledge.

Cohn, E., & Teel, S. J. (1991). Participation in a teacher incentive program and student achievement in reading and math. Economics Working Paper. (ERIC Document Reproduction Service No. ED340709).

Cordes, C. (1983). Research finds little merit in merit pay. *American Psychological Association* Monitor, 14, 10.

Darling-Hammond, L. (1986). A proposal for evaluation in the teaching profession. *Elementary School Journal,* 86, 531–551.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57*(2), 120-138. doi:10.1177/0022487105283796

Dee, T., & 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.

Duval, S., & Tweedie, R. (2000). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89–98.

Eckert, J. (2009). More than widgets: TAP: A systematic approach to increased teaching effectiveness. *National Institute for Excellence in Teaching Reports*. Retrieved from http://www.tapsystem.org/resources/resources.taf?page=ffo_rpts_eckert

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6,* 161–180.

Field, A. P. (2003). The problems of using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2,* 77–96.

Figlio, D. N., & Kenny, L. (2006). Individual teacher incentives and student performance. *National Bureau of Economic Research Working Paper 12627.* Retrieved January 2, 2007, from http://papers.nber.org/papers/w12627

Fryer, R. (2011). *Teacher incentives and student achievement: Evidence from New York City Public Schools* (NBER Working Paper 16850). Cambridge, MA: National Bureau of Economic Research. Retrieved November 19, 2012, from http://www.economics.harvard.edu/faculty/fryer/files/teacher%2Bincentives.pdf

Glazerman, S., Chiang, H., Wellington, A., Constantin, J., Player, D. (2011). Impacts of performance pay under the Teacher Incentive Fund: Study design report. Mathematica Policy Research, Inc. Report (Reference Number 0671.500). Retrieved from http://mathematica-mpr.com/publications/pdfs/education/performpay_TIF.pdf

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. J. (2010). Evaluating teachers: The important role of value-added. *Brookings Brown Center Task Group on Teacher Quality*. Retrieved from http://www.brookings.edu/research/reports/2010/11/17-evaluating-teachers

Glewwe, P., Ilias, N., & Kremer. M. (2003). Teacher incentives. NBER working paper 9671. Retrieved November 19, 2012, from http://www.nber.org/papers/w9671.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 1,* 50-55.

Goldhaber, D., DeArmond, M., Player, D., & Choi, H. (2008). Why do so few public school districts use merit pay? *Journal of Education Finance, 33*(3), 262-289.

Goorian, B. (2000). Alternative teacher compensation: ERIC digest #142. ERIC Clearinghouse on Educational Management. (ERIC Document Reproduction Service No. ED446368).

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology, 149,* 876-883.

Hanushek, E. (2007). The single-salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82*(4), 574-586.

Hanushek, E. (2013). Why educators' wages must be revamped now. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2013/02/06/20hanushek_ep.h32.html

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Harvey-Beavis, O. (2003). Performance-related rewards for teachers: A literature review. Distributed at the Organization for Economic Cooperation and Development (OECD).

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs, *Journal of Educational and Behavioral Statistics, 32*(4), 341-370.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560.

Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Institute of Education Sciences. (n.d.) *What Works Clearinghouse procedures and standards Handbook (3.0)*. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Jacobson, S. L. (1988). The distribution of salary increments and its effect on teacher retention. *Educational Administration Quarterly, 24,* 178–99.

Kelley, C. (1998). The Kentucky school-based performance award program: School-level effects. *Educational Policy, 12,* 305-24.

Kelley, C. (2000). *Douglas County Colorado performance pay plan*. Madison, WI: Consortium for Policy Research in Education.

Kohn, A. (1993). Why incentive plans cannot work. *Harvard Business Review, 71*(5), 54-63.

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

Lavy, V. (2002). "Paying for performance: The effect of financial incentives on teachers' effort and students' scholastic outcomes." Hebrew University Working Paper. Available from the Social Science Research Network. Retrieved from http://www.ssrn.com/

Lavy, V. (2004). Performance pay and teachers' effort, productivity and grading ethics. *National Bureau for Economic Research Working Paper 10622*. Cambridge: NBER. Retrieved from http://www.nber.org/papers/w10622.

Lazear, E. (2002). Performance pay and productivity. *American Economic Review, 90,* 1346-1361.

Lazear, E. P. (1996). Performance pay and productivity. *National Bureau of Economic Research, NBER Working Paper Series #5672*. Retrieved from http://www.nber.org/papers/w5672.pdf

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

Moore-Johnson, S. (1984). Merit pay for teachers: A poor prescription for reform. *Harvard Educational Review, 54*, 175-185.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy, 119*(1), 39-77.

Murnane, R. J., & Cohen, D. K., (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review, 56*, 1–17.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group- randomized trials. *Evaluation Review, 27*, 79-103.

Murray, D. M., Vamell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health, 94*, 423-432.

National Education Association (n.d.). Myths and facts about educator pay. Retrieved from http://www.nea.org/home/12661.htm

National Council on Teacher Quality. (2010). Teacher layoffs: Rethinking “last-hired, first-fired” policies. Retrieved from http://www.nctq.org/p/docs/nctq_dc_layoffs.pdf

National Research Council. (1992). Combining information: Statistical issues and opportunities for research. Washington, DC: National Academy Press.

Odden, A. (2000). New and better forms of teacher compensation are possible. *Phi Delta Kappan, 81*, 361-366.

Odden, A., & Kelley, C. (1997). Paying teachers for what they know and do: New and smarter compensation strategies to improve schools. Thousand Oaks, CA: Corwin Press, Inc.

Odden, A., & Kelley, C. (2002). Paying teachers for what they know and do: New and smarter compensation strategies to improve schools (2nd Ed). Thousand Oaks, CA: Corwin Press, Inc.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Plucker, J. A., Zapf, J. S., & McNabb, S. A. (2005). Rewarding teachers for students’ performance: Improving teaching through alternative teacher compensation programs. *Center for Evaluation & Education Policy: Education Policy Brief, 3(5).*

Podgursky, M.J., & Springer, M.G. (2007). Teacher performance pay: A review. *Journal of*

Policy Analysis and Management, *26*(4), 909-949.

Podgursky, M. (2006). Teams versus bureaucracies: Personnel policy, wage-setting, and teacher quality in traditional public, charter, and private schools. *Education Working Page Archive*, 1-31.

Protsik, J. (1995). History of teacher pay and incentive reforms. *Consortium for Policy Research in Education*. Madison, WI: Finance Center.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education.* New York, NY: Basic Books.

Ramirez, A. (2011, January). Merit pay misfires. Educational Leadership, 68(4), 55-58.

Ritter, G. W., & Barnett, J. H. (2013). *A straightforward guide to merit pay: Encouraging and rewarding schoolwide improvement*. Thousand Oaks, CA: Corwin.

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

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

Sawchuck, S. (2010, June). Merit-pay model pushed by Duncan shows no achievement edge. *Education Week, 29*(33), 1, 21.

Shulman, L. S. (1988). A union of insufficiencies: Strategies for teacher assessment in a period of educational reform. *Educational Leadership, 46*(3), 36-41.

Springer, M., & Winters, M. (2009). New York City's schoolwide bonus pay program: Early *evidence from a randomized trial.* Retrieved from Vanderbilt University, National Center on Performance Incentives. Retrieved from, http://www.performanceincentives.org/data/files/news/PapersNews/200902_SpringerWinters_BonusPayProgram1.pdf

Verma, V., & Lee, T. (1996). An analysis of sampling errors for demographic and health surveys. *International Statistical Review, 64*, 265-294.

Wilson, D. B. (2006). *Meta-analysis macros for SAS, SPSS, and Stata*. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html.

Wilson, D., & Lipsey, M. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*(4), 413-429.

Yuan, K., Le, V., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. Educational Evaluation and Policy Analysis, 35(1), 3-22. doi: 10.3102/0162373712462625

REVIEW AUTHORS

Lead review author:

Name:	**Joshua Barnett**
Title:	Dr.
Affiliation:	
Address:	12839 E Yucca Street
City, State, Province or County:	Scottsdale, Arizona
Postal Code:	85259
Country:	USA
Phone:	+1 (479) 387-8973
Mobile:	
Email:	joshuahbarnett@gmail.com

Co-author(s):

Name:	**Gary Ritter**
Title:	Professor
Affiliation:	University of Arkansas
Address:	207 Graduate Education Building
City, State, Province or County:	Fayetteville, Arkansas
Postal Code:	72701
Country:	USA
Phone:	+1 (479) 575-4971
Mobile:	
Email:	garyr@uark.edu

Name:	Nathan Jensen
Title:	Dr.
Affiliation:	
Address:	121 NW Everett Street
City, State, Province or County:	Portland, Oregon
Postal Code:	97209
Country:	USA
Phone:	+1 (503) 548-5091

Mobile:	
Email:	nate.jensen@nwea.org

Name:	**Wen-Juo Lo**
Title:	Dr.
Affiliation:	University of Arkansas
Address:	247 Graduate Education Building
City, State, Province or County:	Fayetteville, Arkansas
Postal Code:	72701
Country:	USA
Phone:	+1 (479) 575-6321
Mobile:	
Email:	wlo@uark.edu

Name:	**George Denny (deceased)**
Title:	
Affiliation:	
Address:	
City, State, Province or County:	
Postal Code:	
Country:	
Phone:	
Mobile:	
Email:	

ROLES AND RESPONSIBLIITIES

Please give brief description of content and methodological expertise within the review team. The recommended optimal review team composition includes at least one person on the review team who has content expertise, at least one person who has methodological expertise and at least one person who has statistical expertise. It is also recommended to have one person with information retrieval expertise.

Who is responsible for the below areas? Please list their names:

- Content:

Joshua, Gary, and Nathan have all worked for approximately a decade on various merit pay projects. Joshua Barnett has worked in Arkansas, Arizona, and internationally on merit pay issues from building programs to evaluating existing programs. Gary Ritter has worked on merit pay issues in Arkansas, and has also presented evidence on merit pay issues to the state legislatures of Arkansas, Oklahoma, Florida, and to the federal subcommittee on education. Gary has extensive experience designing and evaluating merit pay programs and is well connected to the other leading researchers on this topic. Nathan Jensen has also worked extensively building and evaluating merit pay programs in three separate school districts in Arkansas, including urban, rural, and charter districts, and helped develop a merit pay program for a charter school in Colorado.

- Systematic review methods:

Joshua, Gary, and Nathan have all worked on at least one systematic review of the literature. Joshua and Nathan conducted systematic reviews for their dissertation projects (2007, 2012 respectively). Joshua and Gary also previously worked on a Campbell Collaboration review utilizing systematic review methods. Additionally, Gary has taught courses directly aimed at conducting systematic reviews.

- Statistical analysis:

Wenjuo Lo has worked for a decade as a methodological expert on research projects and will primarily be involved with the analysis section of the review. Wenjuo is an associate professor in the education research and statistics department at the University of Arkansas.

- Information retrieval:

Joshua and Gary have worked previously on a Campbell Collaboration review on volunteer tutoring and are familiar with the process and requirements of retrieving information. Additionally, Gary and Wenjuo have a connection to obtain all articles through their affiliation with the University of Arkansas.

SOURCES OF SUPPORT

This evaluation received support from the Campbell Collaboration's Education Coordinating Group.

DECLARATIONS OF INTEREST

For their dissertations, Joshua Barnett and Nathan Jensen completed impact evaluations of merit based programs that are likely to be reviewed for potential inclusion in the systematic review of the literature. Gary Ritter served as chair for Joshua's dissertation, and as part of

Nathan's committee. George Denny served as a committee member for Joshua. These studies will be thoroughly examined to make certain they fit within the criteria of the review.

The reviewers contend that their experience with merit pay programs will not affect their view of the research. Rather, their exposure and familiarity with programs make their objectivity into understanding how and to what degree programs are effective more tenable. Finally, any potential bias is well counter-balanced by the explicit and transparent methods described throughout the review.

PRELIMINARY TIMEFRAME

The reviewers intend to complete the review by October 2014. As previously noted, the review team has already conducted extensive work in this area of research; therefore, numerous studies anticipated to be included within the review have already been collected and examined. We will again review these studies for accuracy in data extraction and analysis and add additional articles obtained from the search strategy previously described.

The proposed project timeline is as follows:

Date Completed	Milestones/Task
December 5, 2012	• Register title with Campbell Collaboration – initial submission
February 1, 2013	• Submit final title registration, pending revisions from ECG
March 15, 2013	• Submit initial draft of protocol for systematic review, including inclusion and exclusion criteria, search terms and search engines and decisions on unpublished literature.
September 16, 2013	• Submit revised draft of protocol for systematic review, pending revisions from peer reviewers
May 15, 2014	• Submit final protocol for systematic review
August 1, 2014	• Submit initial draft of review
November 15, 2014	• Submit revised draft of review pending revisions from peer reviewers
February 1, 2015	• Submit final review

PLANS FOR UPDATING THE REVIEW

The authors anticipate updating the review on a five year cycle, pending continued public interest and research on the topic and the availability of funding.

AUTHOR DECLARATION

Authors' responsibilities

By completing this form, you accept responsibility for preparing, maintaining and updating the review in accordance with Campbell Collaboration policy. The Campbell Collaboration will provide as much support as possible to assist with the preparation of the review.

A draft review must be submitted to the relevant Coordinating Group within two years of protocol publication. If drafts are not submitted before the agreed deadlines, or if we are unable to contact you for an extended period, the relevant Coordinating Group has the right to de-register the title or transfer the title to alternative authors. The Coordinating Group also has the right to de-register or transfer the title if it does not meet the standards of the Coordinating Group and/or the Campbell Collaboration.

You accept responsibility for maintaining the review in light of new evidence, comments and criticisms, and other developments, and updating the review at least once every five years, or, if requested, transferring responsibility for maintaining the review to others as agreed with the Coordinating Group.

Publication in the Campbell Library

The support of the Coordinating Group in preparing your review is conditional upon your agreement to publish the protocol, finished review, and subsequent updates in the Campbell Library. The Campbell Collaboration places no restrictions on publication of the findings of a Campbell systematic review in a more abbreviated form as a journal article either before or after the publication of the monograph version in *Campbell Systematic Reviews*. Some journals, however, have restrictions that preclude publication of findings that have been, or will be, reported elsewhere and authors considering publication in such a journal should be aware of possible conflict with publication of the monograph version in *Campbell Systematic Reviews*. Publication in a journal after publication or in press status in *Campbell Systematic Reviews* should acknowledge the Campbell version and include a citation to it. Note that systematic reviews published in *Campbell Systematic Reviews* and co-registered with the Cochrane Collaboration may have additional requirements or restrictions for co-publication. Review authors accept responsibility for meeting any co-publication requirements.

I understand the commitment required to undertake a Campbell review, and agree to publish in the Campbell Library. Signed on behalf of the authors:

Form completed by:	**Date:**
Joshua H. Barnett	**April 27, 2014**
Joshua H. Barnett	

APPENDIX A: CODING GUIDE

The following fields will be used to code and extract data from each article. All data will be initially coded in Excel for tracking purposes and warehousing information.

I. Relevance Screening

A. Study Full Citation (APA style)

B. Is the document about an intervention study?

1. Yes
2. Unclear
3. No (explain) – STOP REVIEW

C. Type of publication

1. book
2. journal article
3. book chapter (in an edited book)
4. thesis or dissertation
5. technical report
6. conference paper
7. other:____________________
8. unreported/cannot tell

D. Does the study use an eligible research design?

1. Yes – RCT
2. Yes – QED
3. Yes – other:____________________
4. No – STOP REVIEW

E. Does the study examine a merit pay/performance pay program?

1. Yes
2. No (explain) – STOP REVIEW

F. Does the study include the target population for this review?

1. Yes – Pk-12 students
2. Yes – Pk-12 teachers
3. Yes – Pk12-teachers and students
4. Yes - other:____________________
5. No (explain) – STOP REVIEW

G. Country in which study was conducted

1. USA

2. Canada
3. Great Britain
4. other English speaking
5. other non-English speaking
6. other:____________________
7. unreported/cannot tell

H. Study setting
1. public school
2. private school
3. both
4. other:____________________
5. unreported/cannot tell

I. Study location
1. urban
2. suburban
3. rural
4. other:____________________
5. mixed environment
6. unreported/cannot tell

J. Timing: Was the intervention conducted since 2000?
1. Yes – (year of intervention)?
2. No (explain) – STOP REVIEW

K. RELEVANCY SCREENING:
1. Does the study pass the relevancy screening process?
a) Include an explanation of why study fails, if it does so.

II. Quality Screening
A. How were the individuals assigned to the treatment and comparison groups/how were the groups formed?
B. Describe the counterfactual condition?
C. Are there any notable confounds with the intervention?
D. Attrition:
1. If the study is an RCT, document the level of overall and differential attrition for each data source.
a) If different, calculate the attrition level for the unit of assignment and the unit of analysis
b) Comments
E. Baseline equivalence

1. If the study is a QED or RCT with high attrition, does it meet baseline equivalence standards?

a) Complete table for each outcome measure to determine equivalence?

	Baseline Measure									
Outcome		*Intervention*				*Comparison*				
	x_I	s_I	n_I		x_C	s_C	n_C		*g*	*Equiv?*

F. QUALITY SCREENING – does the study pass the quality screening stage?

1. Include an explanation of why the study fails, if it does so. Summarize any study concerns raised through the screening process.

III. Study Details

A. Describe the intervention program

1. Who is eligible for an award?

2. What is the range/amount of awards?

3. Are the awards individual, group, or combination based (what percentages of each)?

4. How is teacher performance measured (peer observations; principal/administrator observations; student achievement measures; schoolwide measures; other measures)?

B. Describe the demographic composition of the treatment group

C. Describe the demographic composition of the comparison/control group

D. Describe the population in the program:

1. Number of districts

2. Number of schools

3. Number of teachers

4. Number of students

5. Age/grades involved

6. Race/ethnicity of participants

7. Ability/achievement level of participants

8. Other characteristics

E. Program components

1. Additional pay for teachers

a) Individually

b) Grouped

(1) By grade

(2) By content

(3) By school

2. Additional pay for administrators

3. Other:_______________

F. Subgroup analyses explored within the study

G. Outcome Measures

	Name of Outcome Measure	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5	Measure 6	Measure …
Control	Pre							
	Post							
	SD							
	N							
Treatment	Pre							
	Post							
	SD							
	N							
Effect size								
P value								
Reliability								
Type of reliability coefficient								
How was measure administered/collected?								