NORC
at the UNIVERSITY of CHICAGO

NORC WORKING PAPER SERIES

Should Students Be Paid for Achievement?

A Review of the Impact of Monetary Incentives on Test Performance

WP-2015-006 | MAY, 2015

PRESENTED BY:
NORC at the University of Chicago

AUTHORS:
Vi-Nhuan Le

AUTHOR INFORMATION

Vi-Nhuan Le
NORC at the University of Chicago
55 E. Monroe St., 30th Floor
Chicago, IL 60603
Office: 312-357-3865
Fax: 312-759-4004
le-vinhuan@norc.org

Table of Contents

List of Tables

List of Exhibits

Abstract

Despite the proliferation of programs aimed at improving student achievement by offering monetary incentives, there has yet to be a systematic review of the literature that summarizes the potential impact of these incentives on test performance and other correlates of student achievement. Using meta-analytic methods on 15 studies that yielded 18 independent treatment estimates, I found the impact to be weakly positive for overall achievement (i.e., test scores across multiple subjects), as well as for mathematics achievement, but null for reading/language arts achievement. There was no impact of monetary incentives on students' intrinsic motivation, attendance, or self-reported study habits. I supplemented the meta-analysis with a narrative review, where the evidence was mixed with regards to whether treatment estimates could be sustained after the removal of the incentives and whether larger cash payments were associated with stronger program impact. Results are discussed in light of policy implications and future directions for research.

Keywords: financial incentives; student test performance

Although providing incentives to students to attain satisfactory academic performance is a controversial practice, many schools throughout the United States use reward programs as a means of improving academic achievement. Media reports abound with anecdotes of the myriad of incentives that students can receive for reaching specified academic benchmarks, from special recognition at school assemblies to extra school privileges and even iPods (Usher & Kober, 2012). In a survey of 250 charter school principals from 17 states, Raymond (2008) found that 57% of responding principals indicated that they used incentives with their students as a way of raising student achievement. In recent years, cash incentive programs have gained traction, with districts throughout the country paying students for academic achievement. For example, in 2008, the Baltimore City Public School district implemented a monetary incentive program that paid 10th and 11th graders who had previously failed one of their state graduation exams up to $110 if the students improved their scores on benchmark assessments (Ash, 2008). Similarly, 11th graders participating in Atlanta's Learn and Earn program could receive up to $125 if they earned at least a B average in their mathematics and science courses and passed the state exams in those subjects (Ash, 2008).

Despite the prevalence of incentives as a strategy to improve student achievement, until recently little was known about the effectiveness of incentives, especially monetary incentives, on student achievement. Within the past decade, however, there has been a burgeoning body of research in both the domestic and international contexts that has rigorously evaluated the impact of cash incentives on student test performance. Given the amount of resources and attention that policymakers have invested in providing monetary incentives to students as a way to improve education, there is a need for a systematic review that quantifies the impact of incentives on student achievement.

The goal of this study is to use meta-analysis to synthesize the results across the evaluations, and estimate the impact of monetary incentives on student test performance and other correlates of student achievement. I supplement the meta-analysis with a narrative review of findings that did not have a sufficient number of studies to reliably synthesize across studies, but had important ramifications for the design of future cash incentive programs. This study is guided by six research questions shown below. The first two questions are addressed via meta-analysis and the latter four questions are addressed through a narrative review. The research questions include:

1. What is the impact of cash incentives on student test performance and other correlates of student achievement, including intrinsic motivation, attendance, and study habits?

2. Does the impact vary by particular subgroups, including location (i.e., international students versus students in the U.S.), schooling level (i.e., elementary versus secondary grades), initial achievement level, gender, and race/ethnicity?
3. Are the effects of incentives on test performance sustained, once the incentives are removed?
4. Is there a relationship between magnitude of program impact and the size of the monetary incentive?
5. What are the features of promising incentive programs?
6. Are there any unintended consequences of implementing incentive programs?

This manuscript is organized as follows. I begin with an overview of the arguments put forth by critics and supporters of monetary incentive programs. Next, I describe previous literature that has examined the impact of monetary incentives on test performance in laboratory settings. I then describe the analytic approach used in the study, including the search methods, inclusion criteria, and meta-analytic techniques used to estimate effect sizes. This is followed by the results of the meta-analysis, then the results of the narrative review. I conclude with implications of the results for future policy and research.

Expectancy-Value Theory as Rationale for Providing Monetary Incentives to Students

There are a number of theories underlying the premise that monetary incentives can improve student achievement, but the framework adopted by this study is the expectancy value theory of achievement motivation (Atkinson, 1957; Eccles et al., 1983; Wigfield, 1994; Wigfield & Eccles, 1992). Under this theory, students' effort, persistence, and performance on an achievement task are dependent on students' beliefs about their chances of performing well on the task (i.e., the expectancy component) and on the subjective values that they place on the task (i.e., the value component) (Wigfield & Eccles, 1999). Although there are a number of factors that affect students' valuation of an achievement task, variables such as the importance of doing well for one's own sense of identity, the intrinsic interest held by the student when performing the activity, the usefulness of the task for achieving a future goal, and the costs associated with performing the task have been identified as key determinants of a task's value (Eccles et al., 1983; Eccles, O'Neill, & Wigfield, 2005). Expectancy value theorists argue that students may not put forth their best effort on the achievement tasks because they either have low expectancy, low valuation of the achievement tasks, or both.

Studies have suggested that students may have low subjective values for achievement tasks because the costs and effort required to perform well are upfront and high, but the benefits are delayed and may not be readily apparent or understood (Barrow & Rouse, 2013). Monetary rewards for achieving prescribed benchmarks on a test can mitigate these discrepancies by allowing students to more quickly realize the

pay-offs of their hard work in a concrete manner (Wallace, 2009). Thus, monetary incentives can enhance the utility value that students place on performing well on standardized tests. Due to their increased valuation of the task, students may exert more effort to perform well on the tests. Students' increased effort can manifest itself in a number of ways, such as through an increase in school attendance or in the time spent on schoolwork. In the process of pursuing the monetary reward, students may also foster long-lasting, effective study habits or develop higher self-confidence (Wallace, 2009), which can increase students' expectancy of performing well. Thus, the mechanism by which monetary incentives improve student achievement is through an enhancement of students' motivation, which then increases students' effort, which, in turn, leads to greater learning and better test performance.

Monetary Incentives as a Detriment to Intrinsic Motivation

Despite the theoretical appeal of expectancy value theory as a rationale for providing monetary incentives for student achievement, many motivational theorists question the premise that monetary incentives will necessarily increase students' motivation. A meta-analysis conducted by Deci, Koestner, and Ryan (1999) found that providing performance-contingent rewards can actually undermine students' intrinsic motivation, especially if students initially had a high interest in the task. In a classic experiment, Lepper, Greene, and Nisbett (1973) found that providing an external reward to young children to draw and color pictures resulted in a subsequent decrease in children's intrinsic interest in drawing, relative to children who had not received a reward. In a similar vein, Frey and Goette (2000) found that high school volunteers who were collecting donations for charity put forth more effort when they were not compensated than when a small payment was offered. Deci et al. (2001) suggested that tangible rewards decrease intrinsic motivation because the recipients can perceive the reward as an attempt to control their behaviors.

Some motivational theorists argue that even if external rewards could improve performance, the positive impact may be fleeting, as students may withdraw their effort to levels that are lower than initially observed, after the removal of the incentives (Willingham, 2008). In a review of the literature of the relationship between external rewards and intrinsic motivation across a variety of performance tasks, Weinberg (1978) found that after the rewards were withdrawn, the treatment participants were less likely to persist in the activity and expressed lower preference for the activity than the control participants. Similarly, Deci (1971) reported that participants who had been paid to solve a puzzle were less likely to spend time on the puzzle during their free choice period than study participants who were not paid, after the incentive was removed.

Recently, studies have reported neural responses to the removal of financial incentives. Murayama, Matsumoto, Izuma, and Matsumoto (2010) administered a task in which study participants were asked to press a button within five seconds of being prompted by a stopwatch. The task was considered to be a relatively high-interest task because study participants continued to perform the task, even during a free choice period. Half the study participants were not paid for the task, whereas the other half received financial rewards for accuracy (i.e., a winning trial). In a finding that mirrors the Deci (1971) results, after the removal of the financial incentives, the treatment participants engaged in the stopwatch task less frequently than did the control participants, suggesting a decrease in their intrinsic motivation. Notably, there was also a physiological response to the removal of the incentives. During the incentivized session, the treatment participants showed significantly higher blood-oxygen dependent level responses to a winning trial than the control participants. However, once the financial incentives were removed, the treatment participants' blood oxygen-dependent levels decreased to levels lower than the control group. Using an analogous study design, but examining electrophysiological responses, Ma et al. (2014) reported similar results. Namely, participants were less likely to engage in the stopwatch task after the incentive was removed, and declines in participants' intrinsic motivation were accompanied by changes in their electroencephalography recordings in ways that were similar to those observed for the Murayama et al. (2010) study. These studies suggest a neural mechanism by which extrinsic rewards may undermine intrinsic motivation.

Prior Studies on the Impact of Monetary Incentives on Student Test Performance

Until recently, much of what was known about the impact of monetary incentives on student test performance and effort at the K-12 level has been derived primarily from experiments in laboratory settings (Cameron et al., 2005), with the studies reporting mixed results. Baumert and Demmrich (2001) administered a subset of the PISA mathematics tests to nearly 600 German 9th graders. Students in the treatment condition were informed that they would be given 10 Deutsch marks if they could perform better than expected, given their mathematics grade. There were no differences with respect to test performance, effort expended, and motivation for students who were offered a monetary incentive in comparison to students who were not offered an incentive.

In a similar study design, O'Neill et al. (1996) administered a sample of NAEP mathematics to nearly 1,500 8th and 12th graders. Half of the students had been randomly assigned to be paid $1 for each item answered correctly, so that 8th graders could earn up to $41, and 12th graders could earn up to $44. At the 8th grade, there were no differences between the treatment and control participants with respect to metacognition (e.g., self-checking, planning), but there were differences at the 12th grade favoring

treatment students. However, these differences at the 12th grade did not translate to performance differences. There was also a lack of impact of monetary incentives on test scores for the full study sample at the 8th grade, but there was a positive impact for the subset of 8th graders in the treatment group who had remembered the test instructions. In a follow-up study that focused on 12th grade students, O'Neill et al. (2005) attempted to increase the attractiveness of the incentive by increasing the amount paid per item from $1 to $10. The results mirrored their earlier 12th grade findings in that students in the treatment group reported a higher level of effort, but the higher level of effort did not result in higher test scores than students in the control group.

In contrast, Braun et al. (2011) found that monetary incentives could improve 12th graders' NAEP reading performance. Students were either given no incentives, told that they would receive a flat $20 honorarium, or told that they would receive a flat $5 honorarium, but could earn an additional $30, depending on their performance on two randomly selected test questions. In actuality, all students received $35. Students who were told that they would receive a monetary payment reported a higher level of engagement, indicated that they put in more effort on the test, and scored higher than students in the control condition. Furthermore, the impact was stronger for the students whose payments were contingent on their performance.

Taken together, these studies provide important insight to the potential impact of monetary incentives on student test scores, but the laboratory settings circumscribe some of the inferences that can be drawn. For example, students may be highly motivated by the prospect of the receiving a financial incentive for their test performance, but because the motivational effect occurred at the time of the test administration, students could do very little to increase their content knowledge. By way of contrast, the majority of student incentive programs announces the criteria to obtain the incentives well in advance of the test administration, thereby allowing students to take steps to improve their own learning. Thus, it may be the case that the impact of financial incentives on student achievement is underestimated in laboratory settings.

Methods for Research Synthesis

Literature Search Procedures

To locate potential empirical reports, I used the search terms "financial/cash/monetary incentives to students," and "test-based incentives" within six databases representing multiple disciplines: ERIC, PsycInfo, Sociological Abstracts, Dissertation Abstracts, EconLit, and Google Scholar. The sources

covered both peer-reviewed journal articles as well as non-published reports and working papers. I also examined the reference lists of each study to identify other potentially relevant studies, and searched the Social Sciences Citation Index to identify studies that cited seminal research in the field. Although I did not restrict the year of the search, evaluations of student financial incentives is a relatively new topic, so most studies were published within the last decade. The search resulted in an initial sample of 71 studies that needed a high-level review.

Inclusion Criteria

To be included in the meta-analysis, the study had to meet several criteria. First, the study must have been a field experiment as opposed to a laboratory experiment. Second, the incentive programs needed to provide students with cash rewards for their academic performance. This criterion eliminated the majority of conditional cash transfer programs because the primary goal of these programs is to increase school participation, and therefore, they incentivize school enrollment or student attendance as opposed to student achievement. However, conditional cash transfer programs that included an achievement component (e.g., students must maintain a passing grade) were eligible for inclusion in the meta-analysis. Third, the study needed to include test scores as an outcome. This criterion eliminated studies that examined how incentives increased participation in certain programs, such as Advanced Placement (AP) courses. Although increased participation is an important indicator of the effectiveness of incentives, increased participation does not necessarily translate to higher achievement (Davis, 2014; Southern Regional Educational Board, 2010). Fourth, because K-12 incentive studies are in the nascent stages and could benefit from a systematic analysis, I focused on achievement outcomes at the K-12 level. This criterion eliminated all of the postsecondary scholarship studies. Fifth, I eliminated incentive programs within the welfare policy arena because these studies focused on a very narrow segment of the K-12 population, such as teenage parents, that precluded generalizability to the larger student population. Finally, I excluded research briefs, op-ed pieces, and research narratives (e.g., National Research Council, 2011; Slavin, 2010) because they did not present sufficient methodological details to allow me to estimate program impact. However, these sources were used to identify potentially relevant, empirical studies.

This set of inclusion criteria winnowed the initial sample to 18 studies that necessitated a detailed review. I eliminated one study (Raymond, 2008) because it did not estimate the impact of monetary incentives on student achievement. I eliminated yet another study (Ramsey, 2012) because it did not provide sufficient details for me to convert the ANOVA results into the regression estimate metric used in this study. Finally, I eliminated one study (Spencer et al., 2005) because all eligible students in the program received

a cash award, resulting in a lack of a control group. The final analysis was conducted on 15 unique studies.

Coding Study Information

I coded each study for the following information: (a) type of research report (i.e., dissertation, working paper, published report); (b) research design (i.e., randomized or quasi-experimental); (c) structure of the financial incentive program; (d) study location; (e) number of schools in each of the treatment and control groups; (f) student outcomes; (h) independent variables included in the regression models; and (h) regression coefficients and associated standard errors for the treatment and control conditions for each outcome, delineated by subject, gender, racial/ethnic group, and initial achievement level.

In some studies, both unconditional and conditional regression estimates were reported. In those instances, I analyzed the regression coefficients from models that included controls for student- and school-level variables because the inclusion of the control variables are often used to adjust for imbalances between the treatment and control groups (McEwan, 2014), and can reduce the standard error of the estimated treatment estimate (Duflo et al., 2008). However, it should be noted that analysis of the unconditional regression coefficients remained robust to the findings reported below.

Outcomes Examined

I focused on three outcomes in this study: achievement, intrinsic motivation, and effort. For achievement, I report on overall achievement, defined as the treatment estimate pooled across mathematics, reading/language arts, science, and history/geography/social sciences. I also report on impact separately by mathematics and by reading/language arts, where applicable. All studies included in the review used standardized tests (as opposed to teacher-made tests) to assess achievement. Thus, the achievement measures are likely to have adequate reliability and validity.

Of the studies that assessed intrinsic motivation, all used established measures of motivation, including Ryan's (1982) Intrinsic Motivation Inventory or the Academic Self-Regulation Questionnaire, derived from Ryan and Deci's (2000) Self Determination Theory.

Effort was assessed in two ways. First, I examined school attendance. Second, I examined students' study habits, derived from student questionnaires. Although the specific questionnaire items varied across studies, there were commonalities among the constructs assessed, including the amount of time or effort

students reported spending on homework, the frequency that students participated in class, and the frequency with which students engaged in problematic behaviors during class.

Data Integration Methods

The Regression Coefficient as an Effect Size Index

Following the procedures of other studies (e.g., Cooper et al., 2006; McEwan, 2014; Nieminen et al., 2013), I report a standardized regression coefficient as a measure of effect size. Almost all the studies in the review reported a standardized regression coefficient for the impact of financial incentives on test performance and other outcomes. In the few instances that unstandardized regression coefficients were reported instead, I converted the unstandardized regression coefficient estimate to a standardized regression coefficient by dividing the treatment effect and its standard error by the pooled standard deviation of the outcome variable (McEwan, 2014).

To identify outliers in the regression estimates, I conducted the Grubbs test (1950), which calculates a G statistic, defined as the absolute value of the difference between a given estimate and the mean estimate, divided by the standard deviation. If the G statistic exceeds the critical value for a two-tailed test at a 0.05 significance level, the estimate is considered an outlier. To analyze outliers, I followed the procedures of Cooper et al. (2006), and set the value of the outlier to that of its nearest neighbor and recalculated the effect size.

I used the Comprehensive Meta-Analysis software package (Borenstein, Hedges, Higgins, & Rothstein, 2005) to analyze the data using a random-effects model. Due to differences in the incentive structures and in the population in which the samples were drawn, it is unlikely that the effect sizes could be assumed to share a common mean effect size and be derived from the same distribution, with only sampling error contributing to the variation in the observed effect sizes. Thus, a fixed-effect model is not as appropriate as a random-effects model, which assumes that variation in the observed effect sizes stems from both sampling error and from random variance that reflects differences in the way that the programs were implemented.

In pooling the results across studies, it is important to account for the variation in the reliability of each estimate. Following Kim (2011) and Nieminen et al. (2013), each effect size is weighted by the inverse of the squared standard error of the effect size value. Thus, the weighted mean of the standardized regression

coefficients (i.e., $(\bar{\beta})$) and the standard error of the weighted mean (i.e., $SE(\bar{\beta})$) are given in Equation (1) as:

$$\bar{\beta} = \frac{\sum \beta_i / v_i}{\sum 1/v_i} \quad \text{and} \quad SE(\bar{\beta}) = \sqrt{\frac{1}{\sum 1/v_i}} \qquad \text{Equation (1)}$$

where β_i is the standardized regression estimate associated with study i, and v_i is the associated squared standard error for study i. In addition to reporting the average effect, I also report the 95% confidence interval (CI) around the estimate.

A potential challenge with using standardized regression coefficients as an effect size stems from variation in the predictors included in the regression models, which may make it difficult to compare the regression coefficients across studies (Becker & Wu, 2007). For the set of studies included in this review, there was a fair amount of consistency in the predictors used, especially for the studies conducted in the U.S. Across all studies, the regression models included controls for students' prior achievement, gender, and income level/poverty. In the U.S., all studies also included controls for race/ethnicity, and most included controls for English language learner status and specials needs status. None of the regression models included teacher-level variables as predictors, and variation between regression models stemmed mostly from the inclusion (or exclusion) of a limited set of school-level variables. Because all studies included the key student-level characteristics that have been shown to account for much of the variation in test scores (Goldhaber et. al., 1999), and because school-level variables have been shown to account for a very small percentage of test score variation (Cappell & Ippel, 2004), it is likely that the regression coefficients reported for this review are roughly comparable.

Subgroup Analysis

I also examined effect sizes for the achievement outcomes separately by certain subgroups, with the stipulation that there were at least five separate treatment effects to ensure adequate precision (Williams, 2012). In addition to reporting an aggregate effect, I also report effect sizes separately by location, schooling level, initial achievement level, gender, and race/ethnicity. These subgroups were chosen because the literature suggests effect sizes may vary by these factors. For example, many of the incentive programs in international settings were conducted in developing countries, which differs markedly from the U.S. with respect to school resources and social norms for school attendance. In terms of schooling level, it is possible that older children may be more motivated by incentives than younger children because older children may have a better understanding of money and finances. It is also important to examine whether there are differential impact on students of different achievement levels because

previous studies have found that financial rewards can have a positive impact on high-achieving college students, but a negative impact on lower-achieving college students (Leuven et al., 2010).With respect to gender, scholarship studies conducted at the postsecondary level have found that females are more responsive to financial incentives than males (Angrist, Lang, & Oreopoulos, 2009). Finally, it is important to explore the extent to which underrepresented minorities respond to financial incentives because the findings can help guide strategies for closing the achievement gap.

Independent Hypothesis Tests

A key assumption of meta-analysis is that the treatment effect sizes can be treated as independent effects (Kim, 2011). However, as noted by Cooper et al. (2006), it may be difficult to decide what constitutes an independent estimate of effect, especially when there are multiple outcomes or multiple incentive conditions within the same study. For example, multiple outcomes (e.g., mathematics and reading achievement) are not independent because the same set of students contribute information to both outcomes. Similarly, with multiple incentive conditions in a single study, the same control group serves as the basis for comparisons, thereby violating the tenets of independence.

To address the non-independence in treatment effects, I adopted the shifting unit of analysis (Cooper, 1998). With this procedure, effect sizes are combined or left separate, depending on the outcome being studied. For example, to examine the impact of incentive programs on overall academic achievement, the effect sizes for mathematics and language arts/reading from the same study would be averaged into a single effect size. Thus, this study would contribute only one effect size to the analysis. However, to examine the impact of incentive programs on mathematics and language arts/reading, the study would contribute two separate effect sizes to the analyses.

I adopted an analogous approach with studies that have multiple incentive conditions, and require multiple comparisons against one control group. For example, one study examined the impact of incentives provided to individual students, to teams of students, and to teams of students using a tournament format (Blimpo, 2014). In this case, all three treatment incentives were being compared to the same control group, which meant that the three treatment effects were not independent. Following the recommendations of Borenstein et al. (2009), I collapsed data across the different treatment conditions and used this data to compute a single effect size.

Results

Characteristics of the Included Studies

Table 1 shows the characteristics of the incentive programs included in the review. As shown in Table 1, there is a balance of published and unpublished papers in the analysis. The studies are also balanced with respect to location and schooling level. All but three studies used a randomized design. Five studies—Angrist & Lavy, 2009; Jackson, 2010; Jackson, 2014; Kremer et al., 2009; Li et al., 2013 --- did not report treatment estimates separately by subject. Thus, these studies could only be used to estimate the impact of incentives on overall achievement.

Table 1. Characteristics of the Studies Included in the Meta-Analysis

Study Name	Published	Randomized	Location	Grade Levels	Outcomes	Subgroups	Sample Size/Incentive Structure
Angrist & Lavy (2009)	Yes	Yes	Israel	10th-12th	-- Passing rates on a multi-subject Bagrut certification exam	-- Achievement level -- Gender	-- 20 treatment, 20 control -- NIS500 for taking any Bagrut component test -- NIS1,500 for passing the component tests before senior year -- NIS 6,000 for any senior who received a Bagrut -- NIS 10,000 to a student who passes all the achievement milestones
Barrera-Osorio & Filmer (2010)	No	Yes	Cambodia	4th – 6th	-- Mathematics test -- Working memory test -- Effort	-- Achievement level	-- 101 control, 208 treatment -- Scholarships equivalent to $20 for staying enrolled in school, regular school attendance, and maintaining passing grades -- Two treatment conditions: 1) Targeted students based on poverty 2) Targeted students based on merit
Behrman et al. (2012)[1]	No	Yes	Mexico	10th-12th	-- Mathematics test -- Effort	-- Achievement level -- Gender	-- 60 treatment, 28 control --Three treatment conditions: 1) Students only: incentives range from 2500 to 15,000 pesos, depending on the grade level , initial achievement, and the achievement level attained on the incentivized test 2) Teachers only: incentives range from 0 to 750 pesos, depending on the grade level, initial achievement, and the achievement level attained by students on the incentivized test 3) Students, teachers, and administrators: A combination of the incentive structures described above; in addition, incentives are provided based on the average performance of the class (for students) and the average performance of other students in the school (for teachers and administrators)

Study Name	Published	Randomized	Location	Grade Levels	Outcomes	Subgroups	Sample Size/Incentive Structure
Berry (2013)	No	Yes	India	1st-3rd	-- Reading test -- Attendance		-- 8 treatment and control schools -- 100 rupees based on children's individual mastery of literacy objectives -- Incentives could be awarded to the parents, child, or a toy equivalent
Bettinger (2012)	Yes	Yes	Ohio	3rd – 6th	-- Mathematics test -- Reading test -- Social science test -- Science test -- Intrinsic motivation	-- Achievement level -- Gender	-- 4 treatment and control schools -- $15 for scoring at least proficient on each of five subject area tests -- $20 for scoring advanced or proficient on each of five subject area tests
Blimpo (2014)	Yes	Yes	Benin	10th	-- Multi-subject national exam: mathematics, physics and chemistry or English, natural science, history and geography, writing, French, and reading comprehension	-- Achievement level	-- 72 treatment, 28 control --Three treatment conditions: 1) Individual incentives: ranges between 5000 and 20,000 Francs CFA, for a passing and honor score passing performance, respectively 2) Team incentive: ranges between 20,000 and 80,000 Francs CFA, for a passing and honor score passing performance, respectively 3) Tournament: team who ranks among the top 3 average score receives 320,000 Francs CFA
Fryer (2010)[2]	No	Yes	Washington, DC	6th – 8th	-- Mathematics test -- Reading test -- Attendance -- Effort -- Intrinsic motivation	-- Achievement level -- Gender -- Race/ethnicity	-- 17 treatment, 17 control -- Inputs for school-related behaviors that varied by schools, but typically included attendance, behavior, homework, classwork, and wearing a school uniform -- Students were given one point every day for satisfying each of the 5 metrics, for a total of 50 points per payment period -- Students were paid $2 per point

Study Name	Published	Randomized	Location	Grade Levels	Outcomes	Subgroups	Sample Size/Incentive Structure
Fryer (2011)	Yes	Yes	Dallas Chicago New York	2nd 9th 4th, 7th	-- Mathematics test -- Reading test -- Attendance -- Effort -- Intrinsic motivation	-- Achievement level -- Gender -- Race/ethnicity	-- Dallas: 21 treatment, 21 control Chicago: 20 treatment, 20 control New York: 63 treatment, 58 control -- Dallas: Students paid $2 per book and administered a quiz that they must pass at 80% accuracy -- Chicago: Students paid $50 for an A, $35 for a B, and $20 for a C for grades obtained in 5 core subjects (English, mathematics, science, social science, and gym) -- New York (elementary): Students were given $5 for completing interim tests in mathematics and reading, and earned $25 for a perfect score, for up to $250 per year -- New York (secondary): Students were given $10 for completing interim tests in mathematics and reading, and earned $50 for a perfect score, for up to $500 per year
Fryer and Holden (2013)	No	Yes	Houston	5th	-- Mathematics test -- Reading test -- Attendance -- Effort -- Intrinsic motivation	-- Achievement level -- Gender -- Race/ethnicity	-- 25 treatment, 25 control --Students receive $2 per math objective mastered, as indicated by passing a computerized test; students who mastered 200 objectives received a $100 completion bonus with a special certificate -- Parents receive $2 for each objective mastered and $20 per parent-teacher conference attended to discuss child's math progress -- Teachers earn $6 per parent-teacher conference held and up to $10,100 in performance bonuses for student achievement on standardized tests
Holtzman (2010)	No	No	Alabama Arkansas Connecticut Kentucky Massachusetts Virginia	11th-12th	-- Percent scoring above a score of 3 on various AP subject exams		-- 64 treatment, 128 control -- Students receive $100 for each qualifying score, and subsidies for exam fees -- Teachers receive up to several thousand dollars, depending on the number of students who receive a qualifying score

Study Name	Published	Randomized	Location	Grade Levels	Outcomes	Subgroups	Sample Size/Incentive Structure
Jackson (2010)	Yes	No	Texas	11th – 12th	-- Percent scoring above 1100/24 on the SAT/ACT	-- Gender -- Race/ethnicity	-- 40 treatment, 18 control -- AP teachers receive between $100 and $500 for each AP qualifying score earned by a high school junior or senior enrolled in their course -- Teachers can also receive discretionary bonuses of up to $1000 -- Students receive between $100 and $500 for each qualifying score
Jackson (2014)	Yes	No	Texas	11th – 12th	-- Number of AP exams passed across all subjects		-- 58 treatment, 1413 control -- AP teachers receive between $100 and $500 for each AP qualifying score earned by a high school junior or senior enrolled in their course -- Teachers can also receive discretionary bonuses of up to $1000 -- Students receive between $100 and $500 for each qualifying score
Kremer et al. (2009) [3]	Yes	Yes	Kenya	6th	-- Performance on the Kenya Certificate of Primary Education multi-subject test	-- Achievement level	-- 64 treatment, 63 control -- Scholarships awarded to the top 15% of grade 6 girls in the treatment schools -- Scholarship provided winning girls with a grant of US$6.40 (KSh 500) to her school; a grant of US$12.80 (KSh 1,000) to the recipient; and public recognition at a school awards assembly
Levitt et al. (2013)	No	Yes	Chicago	2nd- 8th, 10th	-- Mathematics test -- Reading test	-- Gender	-- 226 treatment and control classes or school-grade combination -- Financial incentives of $10, $20, or non-financial incentive (i.e., trophy) -- Students were paid if they improved on their score from a previous testing session

Study Name	Published	Randomized	Location	Grade Levels	Outcomes	Subgroups	Sample Size/Incentive Structure
Li et al. (2013)	No	Yes	China	3rd – 6th	-- Mathematics and reading test (combined)	-- Achievement level -- Gender	-- 23 treatment, 35 control -- Two treatment conditions: 1) Cash payments for test performance -- 100 RMB was given to the student who achieved the greatest increase in test scores between the baseline test and the evaluation test. Second and third place runners up were promised 50 RMB each 2) Cash payments for test performance + payments for peer tutoring -- Top students from each class served as a tutor. To encourage peer tutoring, the tutor received a cash prize of the same amount as his or her tutee.

[1] I did not include the incentives provided to teachers in my analysis. I also did not analyze the effort measures because the descriptive statistics lacked sample sizes that would have allowed me to compute an effect size.

[2] This study was the working paper version of the published Fryer (2011) study. The Washington, DC results were included in this version of the manuscript, but eliminated from Fryer (2011).

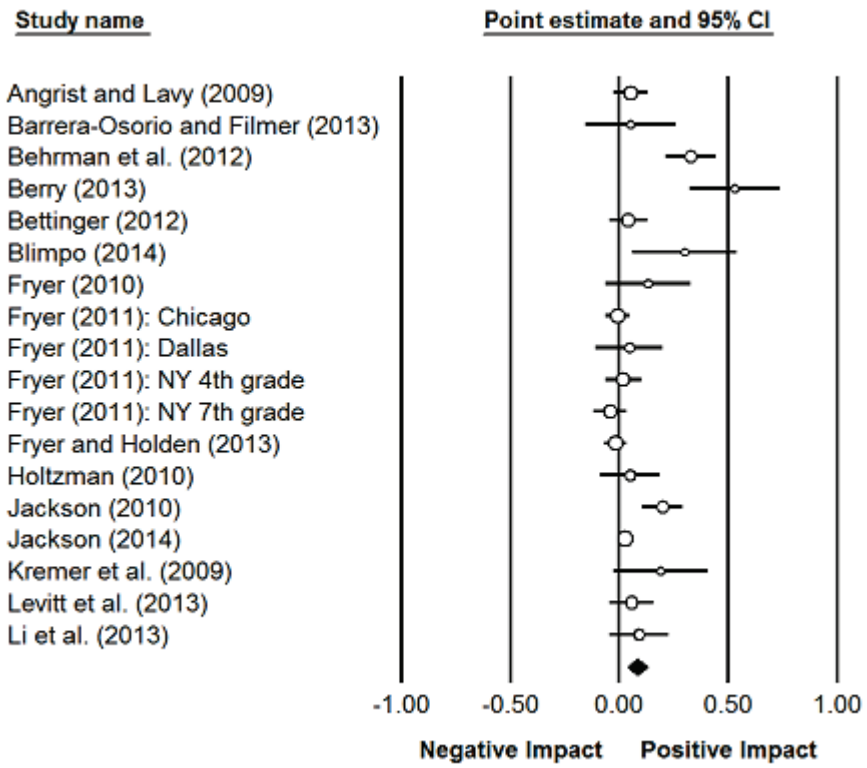
[3] Although the study provided estimates of treatment impact for both boys and girls, I did not analyze the treatment estimates for boys because they were not eligible for the cash incentives.

Fryer (2011) reported on four separate field experiments, which differed with respect to location, grade level, and incentive structures. Because each field experiment had its own distinct treatment and control groups, as well as different achievement measures and incentive structures, the treatment effects are statistically independent. Therefore, I treated the program estimates in the Fryer (2011) study as is if they were derived from four separate studies. This meant the 15 studies in the review yielded 18 treatment estimates.

Impact on Outcomes

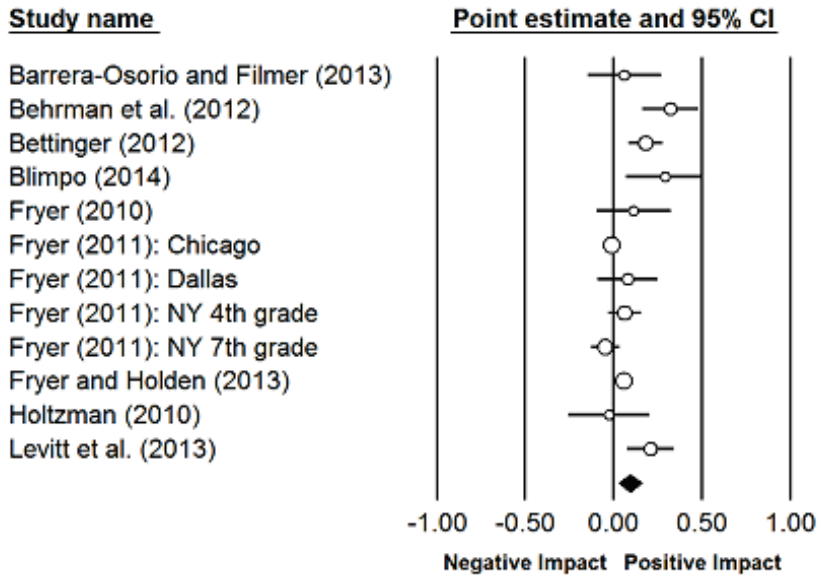
Overall achievement. Figure 1 presents the forest plot showing the impact of monetary incentives on overall achievement. In Figure 1, the impact of each study is denoted by the circle, with the size of the circle representing the influence of the study on the summary estimate. The larger the circle, the larger the influence of the study on the summary estimate. The summary estimate is denoted by the diamond in the last row. As shown in Figure 1, the summary impact across the 18 studies was statistically positive at the 0.01 level of significance ($\bar{\beta}$ = 0.083; 95% CI = 0.040/0.126). I corrected this summary estimate for the Berry (2013) study, which was identified as an outlier, but the summary estimate remained statistically significant at the 0.01 level ($\bar{\beta}$ = 0.074; 95% CI = 0.034/0.114).

Figure 1. Impact of Monetary Incentives on Overall Achievement



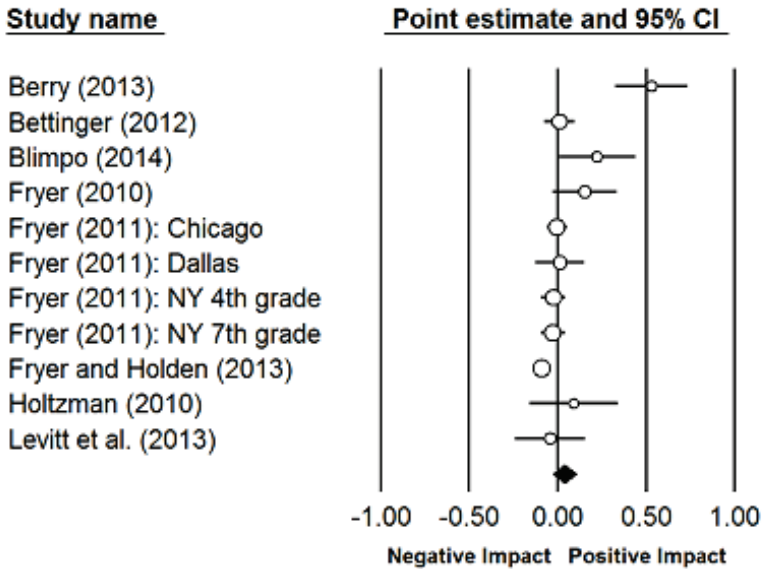
Mathematics. Figure 2 presents the forest plot for mathematics achievement derived from 12 studies. Mirroring the trend for overall achievement, there was a positive impact of monetary incentives on mathematics achievement at the 0.01 level of significance ($\bar{\beta}$ = 0.094; 95% CI = 0.033/0.154).

Figure 2. Impact of Monetary Incentives on Mathematics Achievement



Reading/language arts. Figure 3 presents the forest plot for reading/language arts achievement for 11 studies. Unlike the results presented for overall achievement and for mathematics achievement, incentives did not have an impact on reading/language arts achievement. The summary impact of incentives on reading/language arts achievement, corrected for the Berry (2013) outlying value, was null ($\bar{\beta}$ = 0.006; 95% CI = -0.043/0.055). However, even without this correction, the estimate remained null ($\bar{\beta}$ = 0.037; 95% CI = -0.028/0.102).

Figure 3. Impact of Monetary Incentives on Reading/Language Achievement



Intrinsic motivation, school attendance, and self-reported study habits. The impact of monetary incentives on students' intrinsic motivation was estimated to be null ($\bar{\beta}$ = -0.015; 95% CI = -0.064/0.035; k = 6 effect sizes), as was the impact on school attendance ($\bar{\beta}$ = 0.016; 95% CI = -0.022/0.055; k = 8 effect sizes). In a similar vein, incentives did not impact self-reported study habits ($\bar{\beta}$ = 0.025; 95% CI = -0.015/0.065; k = 6 effect sizes).

Subgroup Analysis

Location. Table 2 provides the results of the subgroup analysis. The impact of cash incentives on overall achievement was stronger for international students ($\bar{\beta}$ = 0.183; 95% CI = 0.074/0.293) than for students in the U.S. ($\bar{\beta}$ = 0.034; 95% CI = 0.001/0.067), although the impact was statistically significant for both groups. In mathematics, there were too few studies to reliably estimate a coefficient for international studies, but a positive impact was observed for studies conducted in the U.S. ($\bar{\beta}$ = 0.061; 95% CI = 0.005/0.117). By way of contrast, in reading/language arts, the impact of incentives was null for U.S. studies ($\bar{\beta}$ = -0.020; 95% CI = -0.062/0.022).

Schooling level. Although there was no impact of financial incentives at either the elementary or secondary grades for reading/language arts achievement, there was a positive impact at both grade levels for overall achievement and for mathematics achievement (see Table 2). For overall achievement, the coefficient estimate at the elementary grades was 0.068 (95% CI = 0.011/0.124) and the estimate at the secondary grades was 0.085 (95% CI = 0.029/0.141). A similar trend was observed for mathematics achievement, where the estimate for the elementary grades was 0.084 (95% CI = 0.034/0.134) and the estimate for the secondary grades was 0.109 (95% CI = 0.014/0.262). For both overall achievement and mathematics achievement, differences between the two schooling levels were not statistically significant, suggesting that incentives were equally effective at both schooling levels.

Initial achievement levels. I compared the impact of financial incentives for lower-achieving students, defined as those whose test performance prior to the implementation of the incentive program was below the median score, against the impact for higher-achieving students, defined as those whose initial test performance was above the median score. Across overall achievement, mathematics achievement, and reading/language arts achievement, the impact of incentives was null for both high- and low-achievers (see Table 2). Thus, there was no evidence that the impact of cash incentives differed by students' initial achievement level.

Gender. In terms of overall achievement, financial incentives had a significant impact for both males ($\bar{\beta}$ = 0.046; 95% CI = 0.004/0.088) and females ($\bar{\beta}$ = 0.048; 95% CI = 0.001/0.096). Furthermore, the positive impact was equally strong for both genders. However, when considering mathematics or reading/language arts separately, the impact on achievement was null for both males and females (see Table 2).

Race/ethnicity. Table 2 shows the impact of financial incentives on the test performance for students of different race/ethnicity. The results show that the impact of incentives did not differ across the various racial/ethnic student groups, nor was the impact significantly different from null for any of the achievement outcomes.

Table 2. Impact of Monetary Incentives on K-12 Test Performance by Subgroup

Subject	*k*	Subgroup	Estimate	95% Lower Confidence Interval	95% Upper Confidence Interval
Overall achievement	7	International	0.183	0.074	0.293
	11	U.S.	0.034	0.001	0.067
Mathematics	3	International[1]		Not estimated	
	9	U.S.	0.061	0.005	0.117
Reading/language arts	2	International[1]		Not estimated	
	9	U.S.	-0.020	-0.062	0.022
Overall achievement	9	Elementary	0.068	0.011	0.124
	10	Secondary	0.085	0.029	0.141
Mathematics	5	Elementary	0.084	0.034	0.134
	6	Secondary	0.109	0.002	0.217
Reading/language arts	5	Elementary	-0.010	-0.082	0.063
	5	Secondary	0.032	-0.040	0.104
Overall achievement	6	Higher-achieving	0.023	-0.011	0.057
	6	Lower-achieving	0.032	-0.022	0.087
Mathematics	6	Higher-achieving	0.043	-0.015	0.101
	6	Lower-achieving	0.056	-0.014	0.125
Reading/language arts	6	Higher-achieving	0.008	-0.026	0.041
	6	Lower-achieving	-0.023	-0.064	0.019
Overall achievement	12	Female	0.048	0.001	0.096
	12	Male	0.046	0.004	0.088
Mathematics	7	Female	0.058	-0.003	0.119
	7	Male	0.027	-0.017	0.071
Reading/language arts	7	Female	-0.022	-0.050	0.006
	7	Male	0.001	-0.035	0.036

Subject	*k*	Subgroup	Estimate	95% Lower Confidence Interval	95% Upper Confidence Interval
Overall achievement	7	Black	0.020	-0.032	0.071
	7	Hispanic	0.009	-0.037	0.055
	7	White	0.044	-0.096	0.184
Mathematics	6	Black	0.012	-0.041	0.065
	6	Hispanic	0.024	-0.041	0.088
	6	White	0.083	-0.221	0.387
Reading/language arts	6	Black	0.007	-0.043	0.056
	6	Hispanic	-0.008	-0.059	0.043
	6	White	-0.037	-0.184	0.111

Sensitivity analyses. Behrman et al. (2012), Fryer and Holden (2013), Holtzman (2014), and Jackson (2010; 2014) incentivized parents and/or teachers in addition to incentivizing students, so the impact of incentives on students was conflated with the impact on other stakeholders. I conducted a sensitivity analysis where the treatment effects were re-estimated without the conflated treatment effects, but effects did not change. Namely, the impact for overall achievement ($\bar{\beta}$ = 0.082; 95% CI = 0.029/0.136) as well for mathematics achievement ($\bar{\beta}$ = 0.109; 95% CI = 0.031/0.187) remained statistically significant at the 0.01 level, whereas the impact for reading/language arts achievement remained null ($\bar{\beta}$ = 0.039; 95% CI = -0.022/0.100).

Although there was a balance of published and non-published works in the review, it is possible that publication bias may nonetheless exist. For example, authors may self-censor, and refrain from putting forth working papers or manuscripts with findings of null or negative impact. I used the Duval and Tweedie's trim-and-fill procedure to account for possible publication bias, and found that the results were robust to this adjustment. One study was trimmed for overall achievement ($\bar{\beta}$ = 0.067; 95% CI = 0.027/0.107), whereas no studies were trimmed for mathematics achievement. Four studies were trimmed for reading/language arts achievement ($\bar{\beta}$ = -0.029; 95% CI = -0.08/0.021). Although the trimmed effect sizes yielded interpretations that were similar to those presented earlier, it is possible that the relatively small number of studies included in the review meant that there was insufficient statistical power to detect bias. Thus, the results should be viewed cautiously.

Narrative Review

I could not reliably synthesize treatment effects for some outcomes because only a few studies examined a particular issue. This was the case as to whether treatment effects persisted after the incentives were removed, and whether there was a relationship between magnitude of program impact and the amount of incentives provided. Similarly, while attempting to identify promising features of incentive programs, I could not pool the results across studies because few studies shared the same incentive structure. However, information relating to these aspects of incentive programs is important because they have implications for the design of future incentive programs. Therefore, instead of a quantitative analysis, I provide a narrative review.

Impact on K-12 Achievement after the Removal of the Incentives

The evidence is mixed as to whether any achievement gains stemming from the implementation of the incentive programs can be sustained after the removal of the incentives. Kremer et al. (2009) found that even one year after the incentive program had ended, the program continued to have a positive impact on test scores, which suggested that the initial learning gains reflected real learning. Similarly, examining the incentive program in Dallas, Fryer (2010) found that a year after the incentives had ended, the treatment group continued to outperform the control group, although the impact was not as strong as when the incentives were in place.

Other studies have found that the achievement gains associated with the monetary incentive programs were short-lived. Bettinger (2012) conducted a multi-year evaluation, where students could be eligible for incentives in one year, but not the next. He found that the achievement gains demonstrated by the incentive recipients in the previous year did not persist into the following year. Levitt et al. (2013) examined test performance during non-incentivized testing sessions. Some of these testing sessions took place a few days after the removal of the incentives, whereas other testing sessions took place a few months after the incentives had ceased. Regardless of the timing of these testing sessions, incentives no longer had a significant impact on subsequent test performance.

Although the Bettinger (2012) and Levitt et al. (2013) studies did not find a lasting impact of cash incentives on test scores, they also did not find that the removal of incentives diminished test performance. That is, the treatment groups' performance did not fall below that of the control groups after the removal of incentives, which would be indicative of a decline in intrinsic motivation. Overall, the test

performance results were consistent with the results from the self-reported questionnaires, both of which indicated that intrinsic motivation was not compromised by the incentive programs.

Impact of the Incentive Programs as a Function of the Size of the Cash Prizes

An important issue to consider is whether the impact of the monetary incentive programs is related to the size of the cash reward. The existing results are contradictory. Fryer and Holden (2013) paid students $2 per mathematics objective mastered. They found that when the amount of incentives was temporarily increased from $2 to $4, and then to $6, the rate of objectives mastered per week also showed a commensurate increase. Namely, when the incentive amount was $2, students mastered an average of 2.05 objectives per week, but when the amount was increased to $4 then to $6, the average number of objectives mastered increased to 3.52 and 5.80, respectively. In a similar vein, Levitt et al. (2013) found that offering a $20 cash prize had a positive effect on test scores, while offering a $10 cash prize had a null effect. However, this effect appeared to be driven mostly by older students, as younger children responded in similar ways to both the larger and smaller incentives.

In contrast to the Fryer and Holden (2013) and Levitt et al. (2013) studies, Jackson (2014) did not find a relationship between program effect size and the size of the reward. The number of AP tests passed was the same for schools that paid $100 per exam as for schools that paid between $101 and $500. Jackson (2010) also found that the impact of monetary incentives on the percent of students scoring above 1100/24 on the SAT/ACT was unrelated to the size of the cash rewards. In fact, incentives of $100 had stronger impact than incentives of $500, with the strongest impact for mid-level incentives (i.e., incentives between $101 and $499). He concluded that there was no evidence to support the notion of a monotonic relationship between magnitude of program impact and size of the cash prize.

Incentive Programs with Multiple Treatment Conditions

Although most studies consisted of a single treatment condition, some studies included two or more treatment conditions, allowing me to directly compare the effectiveness of different types of incentive programs. Blimpo (2014) studied three types of student incentive structures—incentives to individual students, to teams of students, and to teams of students in a tournament format—and found them all to be equally effective. Behrman et al. (2012) also studied the effectiveness of three incentive conditions, which differed with respect to the stakeholders being incentivized: (a) incentives to individual students; (b) incentives to individual teachers; (c) and incentives to students, groups of teachers, and school administrators. They found that providing incentives to individual teachers had no impact on student test

scores, but incentives provided to individual students had a positive impact on test scores. The strongest impact was found when students, groups of teachers, and administrators were all eligible for cash rewards. Notably, teachers in this incentive condition reported spending more outside-of-class time helping students prepare for the exam than teachers in the two other incentive conditions. Although speculative, it is possible that because the rewards were being distributed at the school level, there was more institutional support among the administration to allow the school day to be structured in a way that allowed for more outside-of-class time to prepare students for the test than was the case for the other incentive conditions.

Li et al. (2013) also found that a multipronged incentive structure was most effective. They studied an incentive program in which students who posted the largest achievement gains received a cash prize. In a variation on this incentive structure, they also incentivized peer tutoring in addition to test performance, such that a subset of higher-achieving students were given contracts to tutor other students in the class. If their tutees were among the highest-gaining students, the tutors would receive the same cash prizes as the tutees. Merely paying for test performance had no impact on test scores, but combining incentives for test performance with incentives for peer tutoring showed a positive impact.

Unintended Consequences of Monetary Incentive Programs

One unintended consequence of monetary incentive programs is that it may divert students' attention to the incentivized subjects at the expense of the subjects that are not incentivized. This "substitution effect," however, may depend on initial achievement level. In their study, Fryer and Holden (2013) paid students based on the number of mathematics objectives mastered. High-achieving treatment students mastered more mathematics objectives, scored higher on the standardized mathematics test, and scored comparably on the standardized reading test, relative to high-achieving control students. In contrast, although low-achieving treatment students mastered more mathematics objectives than low-achieving control students, low-achieving treatment students scored comparably to low-achieving control students on the standardized mathematics test, and lower on the standardized reading test. Fryer and Holden (2013) noted that although both high-and low-achieving treatment students put in effort to obtain the prize (as evidenced by the increase in the number of mathematics objectives mastered), this increased effort came at the expense of the low-achieving students' reading performance. Furthermore, even two years after the removal of the incentives, high-achieving treatment students continued to show an advantage in mathematics, without detrimental impact on their reading performance, whereas low-achieving treatment students did not show an advantage in mathematics, and continued to lag behind their low-achieving

control peers in reading. This finding underscores the importance of examining student performance for both incentivized and non-incentivized subjects.

Discussion

The results of this study suggest that financial incentives can improve student achievement. I found a positive impact of monetary incentives on overall achievement and on mathematics achievement, although there was no impact for reading/language arts achievement. This finding is consistent with studies that suggest that incentives may be more effective with concrete subjects (such as mathematics) than with conceptual subjects (such as reading/language arts) (Rouse, 1998). Incentives had a stronger impact for international students than for students in the U.S., although incentives were significantly positive for both groups. I did not find any differences in impact by other subgroups, such as schooling level, initial achievement level, gender, and race/ethnicity.

Incentives did not have an impact on attendance or self-reported study habits, the latter of which is consistent with the findings of Baumert and Demmrich (2001) and O'Neill et al., (1996), both of whom found that incentivizing test performance in laboratory experiments did not lead to higher effort. Notably, regardless of whether intrinsic motivation was operationalized as self-reported questionnaires or as test performance after the removal of the incentives, incentives had no effect on intrinsic motivation.

Study Implications for the Design of Future Incentive Programs

The weak effects found in this study raise questions as to why incentives did not have a stronger impact. One possibility, raised by Fryer (2011), is that offering financial incentives may increase students' motivation to perform well, but students may not know what to do to improve their performance, despite their desire to do so. In interviews with students, Fryer (2011) found that students expressed excitement about the possibility of obtaining a cash prize, but when asked about how they could improve their test performance to attain the reward, students could not readily answer. Students responded with general test-taking strategies (e.g., ensuring that they had correctly read the test item or making sure that their answers were entered correctly), as opposed to strategies that would actually improve their learning (e.g., studying harder, completing their homework, asking teachers for help). Students' lack of understanding about what to do to improve their performance may explain the null impact of incentives on their study habits.

In a similar vein, Li et al. (2013) noted that incentives may help to motivate students, but without being accompanied by additional remediation or supports, incentives, in and of themselves, will not help

students learn the material. This may explain why treatment conditions that incentivized peers or teachers to provide extra assistance to students, such as those implemented by Li et al. (2013) and Behrman et al. (2012), showed stronger effects than treatment conditions that merely paid students for test performance.

Fryer (2011) suggested that incentivizing educational inputs (e.g., reading books) as opposed to outputs (e.g., reaching a performance standard on a test) may lead to stronger effects because inputs allow students to engage in behaviors that can lead to improved performance. By way of contrast, outputs such as "reach proficient level" are abstract goals, and do not offer students concrete steps that will improve their learning. Due to a lack of studies that incentivized inputs as opposed to outputs, I could not explore this hypothesis, but future studies should examine whether cash payments for educational inputs are associated with larger effects.

The results of this study have other implications for the design of future incentive programs. Consistent with the findings from laboratory experiments (O'Neill et al., 2005) as well as the findings from incentive programs conducted at the postsecondary level (Barrows & Rouse, 2013), this study found suggestive evidence that offering a larger cash prize may not necessarily lead to stronger impact than offering a smaller cash prize (Jackson, 2010; 2014). Although the study results do not support the contention that incentive programs undermine intrinsic motivation, the results do suggest that students may engage in substitution, and focus on the incentivized subjects to the detriment of non-incentivized subjects (see Fryer, 2011). This suggests that policymakers may want to design an incentive program that incentivizes multiple subjects, or include stipulations that performance on non-incentivized subjects cannot decline beyond a pre-specified level in order to receive the reward.

Despite the growing knowledge in the area of monetary incentives for students, incentive programs would benefit from additional research that sheds light on how to create maximally effective programs. It is possible that incentives lose their effectiveness if the payments are not immediately provided to students (Levitt et al., 2013; Spencer et al., 2005). It is also possible that some student groups may not be responsive to incentives. Fryer (2011) found that the incentive program in Dallas had a negative impact on the test performance of English language learners, but this finding was not replicated for English language learners participating in the incentive programs in Chicago or New York. More research is needed to determine whether the Dallas findings were anomalous, or whether there were interactions between the particular features of the Dallas incentive program and the characteristics of English language learners that could account for the negative impact. The literature is also sparse with respect to the impact that incentive programs have on students with special needs. Most importantly, future studies

should try to replicate some of the features of promising incentive programs, perhaps through a fractional factorial design that would allow researchers to empirically test multiple features at once.

Study Implications for Educational Practice

The fact that students' test performance could be improved with financial incentives has important implications for the uses and interpretations of the test results. The majority of the achievement measures used in the U.S studies were state achievement tests, which typically do not have any consequences for students. Yet, these same tests can have high-stakes consequences for teachers and schools. Results from state achievement tests, for example, have been used to guide tenure decisions for new teachers, determine bonuses for teachers, and sanction schools for failing to make adequate yearly progress (NRC, 2011). That students may not be putting in maximum effort on the state achievement tests calls into question the validity of using results from these tests as an indicator of quality of teaching or instruction because the test results may not be an accurate reflection of what students have actually learned (Cole & Osterlind, 2008). This suggests that policymakers may need to revisit the practice of using results from the state achievement tests for high-stakes decisions.

A related issue is whether the observed increases in test scores represent real learning, or are instead an artifact of activities that may improve test performance, but not necessarily learning. Studies have shown that when high-stakes consequences are attached to test scores, teachers may narrow the curriculum, and focus on teaching content that is tested, while downplaying content that is untested (Amrein & Berliner, 2012; Hamilton, Stecher, & Yuan, 2012; Yeh, 2005). A consequence of this narrow focus is that test scores may improve, but the gains do not generalize to other tests of similar constructs (Hamilton, Stecher, & Yuan, 2012). Students may react to high-stakes test consequences in similar ways as teachers, such that they may concentrate on studying content that is likely to appear on the test, but at the expense of other content that is less amenable to testing. Indeed, Fryer and Holden (2013) found evidence of substitution effects for students on non-incentivized subjects, and an analogous substitution effect may be operating with respect to specific topics within the incentivized subjects.

If this were the case, then we could expect to see short-lived improvements in test scores, or gains in test scores that are not observed on other achievement measures of the same construct. Thus far, the literature is mixed with respect to whether increases in test scores due to incentives are spurious or not. There have been few follow-up studies that examine achievement after the incentives have been removed, and these studies show mixed results. In a similar vein, there has been no study that has examined whether test score gains observed on the tests used to determine the receipt of the rewards are similarly manifested on

other tests of similar constructs, but do not have high-stakes consequences attached to them. More research is needed to understand the generalizability of test scores gains stemming from financial incentives.

Student-Level Incentives Compared to Other Reforms

An important policy question is the magnitude of student-level financial incentives relative to other interventions. McEwan (in press) conducted a meta-analysis of various school-based interventions designed to improve student achievement in international settings. Relative to other interventions such as class size reduction ($\bar{\beta}$ = 0.117), teacher training ($\bar{\beta}$ = 0.123), or instructional reforms that involve computers or technology ($\bar{\beta}$ = 0.150), student-level financial incentives have a more modest impact ($\bar{\beta}$ = 0.074 pooled across subjects). However, financial incentive programs are relatively inexpensive to implement, especially when compared to other types of reforms that try to improve curriculum, instruction, or teacher quality (Bettinger, 2012; Blimpo, 2014), which may require intensive personnel training or changes to school infrastructure. Fryer (2011) conducted a cost-benefit analysis for the incentive programs included in his study, and found that effect sizes ranging from 0.0006 to 0.016 could have a 5% return on investment. Thus, although the effect sizes for financial incentives are smaller than those for other educational interventions, financial incentives may nonetheless prove to be a more cost-effective strategy for improving achievement than other resource-intensive interventions. Overall, evaluations of whether student-level incentive programs are the best strategy for improving achievement needs to be understood within the larger context of other viable reform options, and should include consideration of the trade-offs between costs and benefits.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Amrein, A. L. &Berliner, D. C. (2012). *An analysis of some unintended and negative consequences of high-stakes testing.* Tempe, AZ: Arizona State University Education Policy Studies Laboratory, Education Policy Research Unit.

* Angrist, J.D.., & Lavy, V. (2009). The effects of high stakes school achievement awards: Evidence from a randomized trial. *American Economic Review, 99*(4)*,* 301–331.

Angrist, J., Lang, D., & Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, *1*(1), 136-163.

Ash, K. (February 13, 2008). Promises of money meant to heighten student motivation. *Education Week*. Retrieved from http://www.edweek.org.

Atkinson, J. W. (1957). Motivational determinants of risk taking behavior. *Psychological Review*, 64, 359–372.

* Barrera-Osorio, F. & Filmer, D. (2013). *Incentivizing schooling for learning: Evidence on the impact of alternative targeting approaches.* Washington, DC: World Bank.

Barrow, L. & Rouse, C.E. (2013). *Financial incentives and educational investment: The impact of performance-based scholarships on student time use.* NBER Working Papers 19351. Chicago, IL: Federal Reserve Bank of Chicago.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16,* 441–462.

* Behrman, J.R., Parker, S.W., Todd, P.E., & Wolpin, K.I. (2013). *Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools.* Philadelphia: Penn Institute for Economic Research at the University of Pennsylvania.

* Berry, James W. 2013. *Child control in education decisions: An evaluation of targeted incentives to learn in India.* Ithaca, New York: Cornell University.

* Bettinger, E.P. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *The Review of Economics and Statistics*, *94*(3), 686–698.

* Blimpo, M.P. (2014). Team incentives for education in developing countries: A randomized field experiment in Benin. *American Economic Journal: Applied Economics, 6(*4*)*, 90–109.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2005). *Comprehensive meta-analysis, version 2*. Englewood, NJ: Biostat.

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

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects on monetary incentives on performance on the 12th-grade NAEP Reading assessment. *Teachers College Record*, *113*(11), 2309-2344.

Cameron, J., Pierce, W.D., Banko, K.M., & Gear, A. (2005). Achievement-based rewards and intrinsic motivation: A test of cognitive mediators. *Journal of Educational Psychology*, *97*(4), 641-655.

Cole, J.S. & Osterlind, S.J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119-130.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., Robinson, J.C., & Patall, E.A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research, 76*(1), 1-62.

Davis, K. (2014, June 27). Enrollment up in AP classes, but testing and scoring lag. *The Baltimore Sun*. Retrieved from http://articles.baltimoresun.com.

Deci, E. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105-115.

Deci, E. L, Koestner, R., & Ryan, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627-668.

Deci, E.L., Koestner, R., & Ryan, R.M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*(1), 1-27.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102*, 1241–1278.

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

Eccles J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Eccles, J.S., O'Neill, S.A., & Wigfield, A. (2005). Ability self-perceptions and subjective task values in adolescents and children. *The Search Institute Series on Developmentally Attentive Community and Society*, *3*, 237-249.

Eccles, J. S., Wigfield, A., Harold, R., & Blumenfeld, P. B. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, *64*, 830–847.

Frey, B. S., & Goette, L. (1999). *Does pay motivate volunteers*? Zürich, CH: Institute for Empirical Research in Economics, Universität Zürich.

* Fryer, R.G. (2010). *Financial incentives and student achievement: Evidence from randomized trials*. Cambridge, MA: Harvard University.

* Fryer, R.G (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, *126*, 1755-1798.

* Fryer, R.G. & Holden, R.T. (2013). *Multitasking, dynamic complementarities, and incentives*: *A cautionary tale*. Cambridge, MA: Harvard University.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199–208.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Journal of the American Statistical Association, 21,* 27–58.

Hamilton, L. S., Stecher, B. M., and Yuan, K. (2012). Standards-based accountability in the United States: Lessons learned and future directions. *Education Inquiry*, *3*(2), 149–170.

* Holtzman, D.J. (2010). *The Advanced Placement Teacher Training Incentive Program (APTIP): Estimating the impact of an incentive and training program on AP taking and passing.* Unpublished manuscript. Retrieved from http://www.socialimpactexchange.org.

* Jackson, C.K. (2010). A little now for a lot later. A look at a Texas Advanced Placement incentive program. *The Journal of Human Resources*, *45*(3), 591-639.

* Jackson, C.K. (2014). Do college-preparatory programs improve long-term outcomes? *Economic Inquiry*, *52*(1), 72-99.

Kim, R.S. (2011). Standardized regression coefficients as indices of effect sizes in meta-analysis. (Unpublished doctoral dissertation). Florida State University, Tallahassee, FL.

* Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, *91*(3), 437 - 456.

Lepper, M., Greene, D., & Nisbett, R. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129-137.

Leuven, E., Oosterbeek, H., & van der Klaauw, B. (2010). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association*, *8*, 1243 – 1265.

* Levitt, S.D., List, J.A., Neckermann, S., & Sado, S. (2013). *The behavioralist goes to school*: *Leveraging behavioral economics to improve educational performance*. Chicago, IL: University of Chicago.

* Li, T., Han, L., Rozelle, S., & Zhang, L. (2010). *Cash incentives, peer tutoring, and parental involvement*: *A study of three educational inputs in a randomized field experiment in China.* Working Paper 221. Stanford, CA: Rural Education Action Project.

Ma, Q., Jin, J., Meng, L., Shen, Q. (2014). The dark side of monetary incentive: how does extrinsic reward crowd out intrinsic motivation. *Neuroreport*, *25*(3), 194-198.

McEwan, P.J. (in press). Improving learning in primary schools of developing countries: A meta- analysis of randomized experiments. *Review of Educational Research*.

Murayama K, et al. (2010) Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceeds of the National Academy of Science of the United States*, *107*, 20911–20916.

Nieminen, P., Lehtiniemi, H., Vähäkangas, K., Huusko, A., & Rautio, A. (2013). Standardised regression coefficient as an effect size index in summarizing findings in epidemiological studies. *Epidemiology Biostatistics and Public Health*, *10*(4), 1-15.

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3,*135–157.

O'Neil, H.F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, *10*(3), 185-208.

Ramsey, S. (2012). *The effect of the Advanced Placement Training and Incentive Program on increasing enrollment and performance of Advanced Placement science exams.* (Unpublished doctoral dissertation). Virginia Commonwealth University, Richmond, VA.

Raymond, M. (2008). *Paying for A's*: *An early exploration of student reward and incentive programs in charter schools*. Stanford, CA: CREDO.

Rouse, C. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics, 113*(2), 553–602.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*, 450-461.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68-78.

Slavin, R.E. (2010). Can financial incentives enhance educational outcomes? Evidence from international experiments. *Educational Research Review*, *5*(1), 68-80.

Southern Regional Educational Board (2010). *Participation and success in the Advanced Placement program continue to grow in SREB states*. Atlanta, GA: Author.

Spencer, M.B., Noll, E., & Cassidy, E. (2005). Monetary incentives in support of academic achievement: Results of a randomized field trial involving high-achieving, low-resource, ethnically diverse urban adolescents. *Evaluation Review*, 29(3), 199-222.

Usher, A. & Kober, N. (2012). *Can money or other rewards motivate students?* Washington, DC: Center on Education Policy.

Wallace, B. D. (2009). Do economic rewards work? *District Administration, 45*(3), 24-27.

Weinberg, R.S. (1978). Relationship between extrinsic reward and intrinsic motivation. *Psychological Reports*, *42*, 1255-1258.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*, 49–78

Wigfield, A., & Eccles, J. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*, 265–310.

Wigfield, A., & Eccles, J. (1999). Expectancy-value theory of motivation. *Contemporary Educational Psychology*, *25*, 68-81.

Williams, R. (2012). *Moderator analyses: Categorical models and meta-regression*. Paper presented at the annual Campbell Collaboration Colloquium, Copenhagen, Denmark.

Willingham, D. T. (2008). Should learning be its own reward? *American Educator, 31*(4), 29-35.

Yeh, S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives*, *13*(43).