Incentives and Test-Based Accountability in Education (2011)

Chapter: 4 Evidence on the Use of Test-Based Incentives

4

Evidence on the Use of Test-Based Incentives

In Chapters 2 and 3, we discuss theory and research on incentives with brief references to tests, and testing with brief references to incentives. In this chapter we delve more fully into the intersection of tests and incentives with the goal of providing an interpretive review of different types of incentives in education in light of the basic research findings about how incentives operate and how they should be evaluated. We focus on rigorous studies that can provide guidance to policy makers about the effects of test-based incentives in education. Although our review does not cover all the available research about the use of test-based incentives in education, we have attempted to include all prominent studies from the past few years that satisfy the criteria we outline below.

In our descriptions of the structure of the test-based incentive programs, we provide information about the key elements that should be considered in designing incentive systems (see Chapter 2): who receives incentives (the targets of the incentive), what performance measures are used, what consequences are attached, and whether supports for improvement are provided. Unfortunately, the available program information often fails to adequately address these elements, which limits our ability to draw inferences about how they affect the outcomes.

In describing evidence about the effects of the incentive programs, we provide information about relevant outcomes other than the tests that are attached to the incentives in order to reduce the likelihood that our conclusions are biased by any distortion that the incentives may cause. We also offer information about changes on high-stakes tests, if it is available, but our focus is on evidence from other measures of the same domain, including both the results of low-stakes tests and other outcomes, such as graduation.

Tables 4-1A, 4-1B, 4-2, and 4-3, presented at the end of the chapter, summarize the descriptive and outcome information discussed in the text below. The studies or groups of studies are referred to below and in the tables as examples; by number, and in some cases additional by letter designations. In both the text and tables, we divide the studies we analyzed into three categories that are familiar to education policy makers and researchers: school-level policies related to the No Child Left Behind (NCLB) Act and its predecessors; high school exit exams; and experiments with teachers and students that use rewards, such as performance pay. Note that the first two categories address policies rather than experiments and so involve larger numbers of students, teachers, and schools and longer implementation periods, but they also present greater difficulties in identifying appropriate comparison groups. NCLB, as the one federal policy discussed in our review, involves particularly difficult challenges in identifying a comparison group.

STUDIES INCLUDED AND FEATURES CONSIDERED

Criteria for Inclusion

Our literature review is limited to studies that allow us to draw causal conclusions about the overall effects of incentive policies and programs.[1] In some cases, programs were planned to include untreated control groups for comparison; in other cases, researchers have carefully documented how to make appropriate comparisons. Because our purpose is to draw causal conclusions about the overall effects of test-based incentives, we exclude several kinds of studies that do not permit such conclusions:

- studies that omit a comparison group, including the evaluations of NCLB carried out by the U.S. Department of Education (Stulich et al., 2007), the Center on Education Policy (2008), and the Northwest Evaluation Association (Cronin et al., 2005), in addition to various well-known earlier studies (e.g., Klein et al., 2000; Richards and Sheu, 1992);
- cross-sectional studies that compare results with and without incentive programs but with no controls for selection into the incentive programs, including well-known studies of exit exams (e.g., Jacob, 2001) and teacher performance pay (e.g., Figlio and Kenny, 2007); and
- studies that focus on contrasting results for students, teachers, or schools that are immediately above or below the threshold for receiving the consequences of the incentive programs,[2] including well-known studies of exit exams (e.g., Martorell, 2004; Papay et al., 2010; Reardon et al., 2010) and school incentives (e.g., Ladd and Lauen, 2009; Reback, 2008; Rouse et al., 2007).

Finally, we exclude programs using incentives that are too new to have meaningful results (e.g., Kemple, 2011; Springer and Winters, 2009).[3] Particularly in the area of performance pay for teachers, there has been strong recent interest in developing new incentive programs, and we expect these will make important additions to the research base in the near future.[4]

Policy and Program Features and Outcomes Considered

The features related to the structure of the incentive programs that we selected for our analysis are derived from four of the five key elements that should be considered in designing incentive programs (see Chapter 2).

[1] For literature reviews that cover a broader range of related studies, see Figlio and Loeb (2010) on school accountability, Podgursky and Springer (2006) on teacher performance pay, and Holme et al. (2010) on high school exit examinations.

Target Our analysis primarily included studies with incentives that were given to schools, teachers, or students, though one case provides an example of incentives given to both students and parents. We coded performance pay programs for teachers as being received by teachers either individually or as a group (Teachers-I or Teachers-G), depending on whether the incentives were based on the performance of each teacher's own students or on the performance of all students in the school.

Performance Measures We used the limited information about the performance measures to code two different features related to the coverage of the measures across subjects and within subjects. For most of the incentive programs we reviewed, the performance measures included only tests, but we noted other measures if they were used. We coded the content coverage across subjects as either narrow or broad, depending on whether the tests included only a portion of the curriculum or most subjects. Usually programs with narrow coverage across subjects focused on language and mathematics tests. When the studies compared results across states where some states used performance measures with broad coverage across subjects and others used performance measures with narrow coverage across subjects, we coded the coverage across subjects as mixed. We also coded the content coverage of the performance measures within subjects as either narrow or broad, depending on whether the test and the performance indicator were sensitive to the full range of content and skill within the subject or to only a portion of the content and skill. For the tests, we looked for information that the tests covered higher-order thinking skills within the subject area. For the performance indicator, we looked for information that the indicator reflected gains across the entire distribution of performance, such as by using a score average or a measure of test score gains rather than a performance level. We coded the coverage of the performance measure within subjects as broad only if both the test and the performance indicator were sensitive to the full range of content and skill.[5]

Consequences With respect to the basic structure of the programs, we coded whether they were focused primarily on penalizing poor performance with sanctions or rewarding performance that meets or exceeds expectations. In the text, we also describe the nature of the consequences and any available information about their frequency, but we did not attempt to code the consequences as large or small because we lacked an objective way of making such a determination.

[2] Such regression discontinuity studies provide interesting causal information about the effect of being above or below the threshold, but they do not provide information about the overall effect of implementing an incentives program.

[3] New York City has recently implemented a performance pay program for teachers in about 200 schools using random assignment of eligible schools (see Springer and Winters, 2009). An initial analysis showed small and negative effects of the program on the tests linked to the incentives, but none of the effects was statistically significant, and the initial analysis used tests that were given less than 3 months after the program was instituted. In addition, New York City's reform effort since mayoral control of the schools began in 2002 includes a schoolwide performance bonus plan that began in the 2007-2008 school year. Initial analysis suggests that scores on the tests attached to the incentives increased faster during the reform period than occurred in comparable urban districts in New York (Kemple, 2011).

[4] See, for example, the various reports on the Texas performance pay program available from the National Center on Performance Incentives (see http://www.performanceincentives.org [June 2011]).

[5] It was often easier to obtain information from the studies about the breadth of the performance indicators than it was to obtain information about the breadth of the tests. Since we required both the test and the indicator to be broad in order to code a program as using a broad performance measure within subjects, we were able to code many programs as using a narrow performance measure within subjects by looking at the performance indicator alone, without needing to obtain information about the test.

Supports To see whether the incentives program takes account of the ability of people to influence their performance, we coded whether or not resources or supports were provided to aid in the attainment of performance goals as part of the incentives program.

Our coding of the incentives structure captures the types of contrasts reflected in the economics literature, but it does not reflect those in the psychology literature about the way that incentives are framed and communicated. In the experimental work discussed in Chapter 2, the contrast between different conditions sometimes involved subtle differences in wording. It is plausible that most of the incentive programs we discuss could have been presented in ways that were either more positive or more negative, depending on whether those in leadership positions characterized them as supporting a shared commitment to learning or as posing an additional burden in already difficult circumstances. Even the contrast between sanctions and rewards fails to measure the way incentives were communicated in a district, school or classroom, since a skillful leader could have described potential sanctions as reaffirming a shared commitment to learning, and an inept leader could have described potential rewards as an attempt to impose external control. In many situations, the contrast between emphasizing one message or the other is subtle—just as it was in the experiments discussed in Chapter 2. The lack of a good measure of the way incentives are framed and communicated is an important limitation in our description of the structure of the different incentive programs.

The features in Table 4-1B related to the outcomes of the incentive programs reflect the importance of providing outcome measures other than the tests that are attached to the incentives. In addition, we looked for information about whether the program effects were distributed across all content areas included in the program and whether they differed for the relatively low- or high-performing students. Our analysis included the following features:

- effect on high-stakes test: the effect of the incentives program on the tests that were attached to the incentives in the program;
- effect on low-stakes test: the effect on tests that were in the same subjects as the tests attached to the program's incentives but that were not themselves attached to those incentives;
- effect on other subject tests: the effect of the program on tests in subjects other than those that were attached to the program's incentives;
- effect on graduation or certification: the effects of the program on graduation or college-bound certification;
- effect on lower performing students: the statistically significant effects of the program for students in the lower half of the achievement distribution; and
- effect on higher performing students: the statistically significant effects of the program for students in the upper half of the achievement distribution.

In the tables, the outcomes columns summarize the outcomes as positive, negative, or not statistically significantly different from zero.[6] If a study provided multiple results, the discussion below and the table entries summarize the overall tendency of the outcomes; if the results diverged, the multiple outcomes are discussed and shown in order of prevalence.

As with our coding of the structural features of the incentive programs, our coding of the outcomes of the programs failed to capture the important outcome from the psychology literature related to changes in dispositions. In general, the studies we analyzed did not provide information about such outcomes; however, a few studies were exceptions to this finding, and for these studies we note their findings related to changes in dispositions in the text.

NCLB AND ITS PREDECESSORS

We identified causal studies related to three examples of school incentives that are in the NCLB mold. Two related to the overall adoption of school incentives across the United States: Example 1 reflects the initiatives in a number of individual states before NCLB, and Example 2 reflects the changes that came with NCLB. Example 3 is Chicago, for both the initial district-level incentives in the 1990s and the implementation of the succeeding NCLB incentives.

Examples 1 and 2: Nationwide School Incentives

A number of states instituted test-based incentives during the 1990s, with consequences for schools that anticipated the consequences that were implemented for all states in 2001 under NCLB (Dee and Jacob, 2007; Hanushek and Raymond, 2005). Under NCLB, schools that do not show adequate yearly progress face escalating consequences. The structure of NCLB defines consequences for schools that involve increasing levels of state intervention and support to bring about improvement. The initial requirements are to file improvement plans, make curriculum changes, and offer their students school choice or tutoring; if progress does not improve as specified, they are required to restructure in various ways. The consequences are based on state tests in reading and mathematics that use state-defined targets for student proficiency. During 2006-2009, the proportion of schools failing to show adequate yearly progress ranged from 29 to 35 percent (Center on Education Policy, 2010). There is mixed information about the implementation of the consequences prescribed under NCLB, with frequent focus on making curriculum and instructional changes, but fewer cases of implementing effective school choice or tutoring options that students use (Center on Education Policy, 2006a).

We treated the incentive programs adopted by many states in the 1990s as roughly similar to NCLB although there were many variations in the incentive structures in the states that may have affected results. For example, North Carolina's school incentives, which were implemented in 1996 and continued alongside NCLB after 2001, are based on test score gains rather than proficiency levels and so are targeted to a broad range of performance rather than a narrow range near the proficiency cut point. Under the two different performance criteria, there were different outcomes: schools facing sanctions under NCLB improved the test scores of lower performing students, while schools facing sanctions under the state program improved the test scores of both lower and higher performing students (Ladd and Lauen, 2009). Unfortunately, there were no studies available that would have allowed us to contrast the overall effect of state incentive programs predating NCLB by the committee's key elements in incentive structure.

We considered three studies that identified causal effects of school incentive policies by comparing changes in states that did and did not use those policies. The studies used the National Assessment of Educational Progress (NAEP) to measure achievement in reading and mathematics for fourth and eighth grade students. For the early period, we used a meta-analysis of 14 studies that compared states that started test-based incentives before NCLB with states that did not (Lee, 2008). For the later period, we used two studies that each performed a complementary analysis that compared states that started using school incentives under NCLB to states that already had school incentives before NCLB (Dee and Jacob, 2009; Wong, Cook, and Steiner, 2009).

Example 1: Pre-NCLB Nationwide School Incentives

For the early period, the meta-analysis by Lee (2008) identified 14 studies that compared results across states with different test-based accountability policies. Most of the studies used longitudinal NAEP data

[6] We used the most lenient level of statistical significance provided in each study, generally p < 0.10 or p < 0.05.

from the 1990s to compare states with different levels of test-based school accountability policy.[7] The studies defined the policy contrasts in a variety of ways and used a variety of analytic strategies. Some of the studies focused on mathematics, and others looked at both mathematics and reading. Most of the studies looked at test results in grades 4 and 8. Across the 76 effect sizes that were calculated from the studies, the average effect size associated with a contrast between states with and without test-based accountability was 0.08 standard deviations (Lee, 2008, p. 625); 66 were positive, 2 were zero, and 8 were negative (pp. 631-638).[8] The study did not report how many of these effects were statistically significant. The meta-analysis did not find significant differences in effect sizes between school and student incentive policies (p. 616), between mathematics and reading (p. 619), between different grade levels (p. 619), or between different racial and ethnic groups (p. 621).

Example 2A: NCLB Nationwide School Incentives (Dee and Jacob)

For the NCLB period, Dee and Jacob (forthcoming) estimated that the imposition of the NCLB requirements in states that had not yet adopted school incentives increased achievement by 2007 in fourth grade mathematics by 7.2 points in the preferred model (Dee and Jacob, forthcoming, Table 3, Panel B). This increase corresponds to an effect size of 0.23 standard deviations. The effects on eighth grade mathematics and fourth grade reading were positive, and the effect on eighth grade reading was negative; none of these other effects was statistically significant.[9] The paper did not provide a formal test of the statistical significance of the subject or grade differences in the effect sizes. Over four combinations of

[7] Given this generalization, the multiple studies in Lee (2008) can be thought of as effectively providing multiple analyses of a single big experiment across states in the 1990s, using different ways of analyzing the available NAEP data. Note that four studies included in Lee (2008) do not fit the generalization in the text: two involve cross-sectional comparisons (Bishop et al., 2001; Lee, 1998) and two focus exclusively on high school exit requirements that are based on minimum competency testing rather than school accountability (Fredericksen, 1994; Jacob, 2001), with one (Jacob, 2001) using the National Education Longitudinal Study rather than NAEP.

[8] The effect sizes are calculated in Lee (2008) from information provided in the original papers. The figure reported in the text is for effect sizes calculated in terms of the standard deviation of student scores. Note that many of the effect sizes reported in the paper are based on the standard deviation of state scores and so are not comparable to the versions calculated in terms of the standard deviation of student scores.

[9] The study notes uncertainty about the reading estimates because the fourth grade data do not follow the linear trend that the statistical model assumes and because the eighth grade data include only two pre-NCLB observations. The results for eighth grade reading were reported only in an appendix.

subject and grade, the average effect size was 0.08 standard deviations.[10] The increase for fourth grade mathematics occurred for both lower and higher performing students (Table 5). Finally, a check for changes in NAEP science test scores showed no effect of NCLB in either fourth or eighth grade on a subject without incentives (Table C4, Panel B), with a small positive effect in grade 4 and a small negative effect in grade 8, neither of which was statistically significant.

Example 2B: NCLB Nationwide School Incentives, Public Schools (Wong, Cook, and Steiner)

Wong, Cook, and Steiner (2009) found similar results for the NCLB period for public schools, though with some differences in their approach. In addition to the contrast between states with and without school incentives before NCLB used by Dee and Jacob, they added a contrast between states with high and low standards. Although high standards did not appear to interact with incentives,[11] the results suggested that the separate effects of the two policies combined in grade 4 reading to produce a statistically significant change.

Across three combinations of subject and grade, the average effect size associated with incentives was 0.12 (Wong, Cook, and Steiner, 2009, Table 14).[12] The effect size was statistically significant only for fourth grade mathematics (Table 13). The paper omitted eighth grade reading, the one test for which Dee and Jacob found negative effects.

[10] We computed the average from the coefficients on the "Total effect by 2007" line of Table 3 in Dee and Jacob (forthcoming) dividing each by the standard deviation of the scores for the different tests provided at the bottom of the table. The results for eighth grade reading were taken from the corresponding line of appendix Table C2. Despite the authors' uncertainty about the reading estimates (see fn. 9), our analysis included them in the overall average in order to provide the best available average of the effect of NCLB that reflects a balance across subjects and grades. When the subjects were considered separately, the average effect for mathematics was 0.17 standard deviations, and the average effect for reading was 0.00 standard deviations.

[11] In the case of fourth grade mathematics, in one specification there was an interaction effect of standards and incentives with borderline statistical significance that suggests that either high standards or incentives alone produced the same effect as the two policies together (Wong, Cook, and Steiner, 2009, Table 13).

[12] We averaged the effect sizes in the "Diff. in Total Δ (2007 or 2009) CA" line of Table 14 of Wong, Cook, and Steiner (2009).

Example 2C: NCLB Nationwide School Incentives, Public and Private Schools (Wong, Cook, and Steiner)

Wong, Cook, and Steiner (2009) also used a comparison between public and private (mostly Catholic) schools as a way to estimate the effects of NCLB, though Dee and Jacob rejected this approach because of the decline in Catholic school enrollment that occurred around the start of NCLB (because of the sex abuse scandal). In addition to comparing public and Catholic schools, the study also compared public and non-Catholic private schools. Over six combinations of subject, grade, and private school type, there was an average effect size of 0.22 standard deviations associated with the change in public school NAEP scores by 2007 or 2009.[13] Although all of the effect sizes were positive, the only one that was marginally significant was for fourth grade mathematics for Catholic private schools (Wong, Cook, and Steiner, 2009, Table 6).

Related Studies About School Incentives

There have been a number of studies of the instructional changes that have accompanied the implementation of school incentives (e.g., Center on Education Policy, 2007a; Hamilton et al., 2007; Rouse et al., 2007; Stecher, 2002; White and Rosenbaum, 2008). In general, these studies found shifts in instruction that were consistent with the performance measures that were attached to the incentives. Some of these changes were aimed at improving achievement broadly, such as increasing total instruction time, improving the alignment of instruction with standards, or adding professional development for teachers. Other changes were focused on the specific structure of the incentives system, such as shifting instruction to focus on aspects that count in the system and away from aspects that do not count: these changes involved an increased focus on tested subjects, on lower performing students at the threshold of attaining proficiency, and on material that closely mimics the tests. These findings about instructional shifts underline the necessity of evaluating the effect of incentives with information from low-stakes tests in the same subjects as the tests attached to incentives, on students at different performance levels, and on subjects not attached to incentives.

In addition to changes in instruction in the subject area, there is evidence of attempts to increase scores in ways that are completely unrelated to improving learning. The attempts included teaching test-taking skills, excluding low-performing students from tests, feeding students high-

[13] We averaged the effect sizes in the "Diff. in Total Δ (2007 or 2009)" lines of Table 7 of Wong, Cook, and Steiner (2009) for the "Public vs. Catholic (Main NAEP)" and "Public vs. Non-Catholic (Main NAEP)" sections of the table.

calorie meals on testing days, providing help to students during a test, and even changing student answers on tests after they were finished (e.g., Cullen and Reback, 2006; Figlio and Getzler, 2006; Figlio and Winicki, 2005; Jacob and Levitt, 2003; Stecher, 2002). The evidence about behaviors that were likely to distort test results again underlines the importance of evaluating the effects of incentives using measures of the same domain that are different than the results of the tests attached to the incentives. It is also important to note, however, that some of the changes that can distort high-stakes tests—such as a focus on the portions of the subject that are easy to test—can also distort low-stakes tests.

Example 3: Chicago School Incentives

The incentives that Chicago introduced in 1996 included sanctions for both schools and students (Jacob, 2005). The school sanctions involved the possibility of reconstituting schools with a high percentage of low-performing students. The student sanctions involved mandatory summer school and retention for students unable to pass exams in the third, sixth, and eighth grades. If students were unable to pass the exams after summer school, they had an additional opportunity to rejoin their class if they could pass the exams in January of the following year. During the first 3 years of the program, retention rates in these grades increased to 10-20 percent, far above the prior level of 1-2 percent (Jacob and Lefgren, 2009).

Jacob (2005) used longitudinal data for Chicago that included the period before the policy took effect and controls for both prior test trends and changes in student demographics. For the 4 years after the start of school incentives, scores on the high-stakes tests in the three grades had increased above predicted trends by about 0.2 standard deviations in reading and 0.3 standard deviations in mathematics (Jacob, 2005, Table 1). Similar results were obtained by comparing the change in Chicago's test score trends when incentives were introduced with the test score trends in other large, midwestern cities (Table 2). Looking across students, there were generally positive effects for both lower and higher performing students in mathematics; for reading, the effects occurred primarily for lower performing students (Table 3). In the lowest decile of students, however, there was some indication that incentives decreased performance. Neal and Schanzenbach (2010) obtained similar results on the distribution of effects across students.

Jacob (2005, Table 5) replicated a version of his analysis with data on low-stakes tests in reading and mathematics. The analysis showed an effect of about 0.2 standard deviations in both subjects 2 years after implementation, but only for the eighth grade; the effect on the low-stakes tests for the third and sixth grade was either negative or was small and not statistically significant. Over nine combinations of subject, grade, and model specification, the average effect size was 0.04. Five of the effects were statistically significant, three of them positive and two of them negative; for the four effects that were not statistically significant, two were positive and two were negative.[14] A direct contrast of the results in mathematics across the three grades showed an average effect size of 0.11 standard deviations on the test attached to the incentive, in comparison with an effect size of 0.04 standard deviations on the test not attached to the incentive. In both cases two of the three effects were statistically significant, but for the high-stakes test both of the significant effects were positive, and for the low-stakes test one was positive and the other was negative.[15]

Jacob (2005, Table 8) also looked at changes in low-stakes tests in science and social studies for students in the fourth and eighth grades, finding that scores in these subjects increased after incentives were introduced. Although the increase in test scores for science and social studies was smaller than for reading and mathematics and occurred primarily with higher performing students, it was positive and so does not suggest a tradeoff between the high-stakes and low-stakes subjects.

HIGH SCHOOL EXIT EXAMS

Use of exit exams has been growing over the past three decades and now includes 25 states and two-thirds of public high school students (Center on Education Policy, 2007b; Warren et al., 2006). There is important variation across states in the nature of the tests used, with general movement over time from minimum competency tests of basic skills below the high school level, to standards-based tests at the ninth and tenth grade levels, to end-of-course tests that are focused on the content of specific high school courses. Exit exams typically involve tests in multiple subjects, all of which must be passed, though many states provide alternate paths that can be substituted for a failure on one or more subject tests (Center on Education Policy, 2006b). States and districts provide a variety of remediation programs and materials for students, as well as assistance to teachers to help prepare students for the exams (Center on Education Policy, 2007b). We identified three causal studies across a large number of states; they used the staggered implementation of exit exams to examine their effect on several different outcomes.

Example 4A: Effects on Achievement

Study 4A looked at long-term trend NAEP results in reading and mathematics for eighth and twelfth grades from 1971 to 2004: it found no effect of the introduction of high school exit exams for either lower or higher performing students (Grodsky et al., 2009, Tables 3-4). Over four combinations of subject and grade, the average effect size was 0.00 standard deviations, evenly divided between small positive and negative effects, and none was statistically significant.[16]

Examples 4B and 4C: Effects on Graduate Rates

Two studies looked at effects on graduation rates. Study 4B used state graduation rates from 1975 to 2002: it found that states adopting more difficult exit exams showed a statistically significant decrease in graduation rates of 2.1 percentage points (Warren et al., 2006, Table 2).[17] This result came from an analysis using Common Core Data that distinguished a high school diploma from a GED (general education development) certificate. An alternate analysis based on census data that used a graduation measure that combined high school diplomas and GED certificates showed no effect of exit exams: this result suggests that the requirement may shift some students from a obtaining a diploma to obtaining a GED.[18]

Study 4C used individual census data for 2000 with state fixed effects that identified changes resulting from exit exam requirements: it found that the requirements for more difficult exams were associated with a decrease in high school graduation—including both diplomas and GED certificates—of about 0.6 percentage points (Dee and Jacob, 2007, Table 6-2).[19] Over three different model specifications, all estimates were negative, and two of them were statistically significant. For the less difficult exit exams, Dee and Jacob (2007) found an average decrease of 0.3 percentage points, with only one of the three estimates statistically significant.

[14] We averaged the estimates for the Illinois Goal Assessment Program (IGAP) test from Table 5 in Jacob (2005), using the models that included controls and prior trends. We did not use the models without controls and prior trends because the study used observational data that cannot support a causal interpretation.

[15] We averaged the estimates for the ITBS and IGAP tests, respectively, in Panel A of Table in Jacob (2005), using the models that included controls and prior trends.

[16] We used the coefficients in the “HSEE” line of Table 3 of Grodsky et al. (2009, Table 2), dividing each by the standard deviation

for reading and math scores, respectively.

[17] We used the estimates based on the Common Core Data with the model that distinguishes between minimum competency and more difficult exit exams (Warren et al., 2006, Table 2).

[18] Outcomes for high school graduates with a regular diploma are substantially better than those with a GED so it is better to distinguish the two outcomes (National Research Council and National Academy of Education, 2011).

[19] We averaged the three estimates in the "More difficult exit exam" line of Table 6-2 of Dee and Jacob (2007) for columns (3), (4), and (5).

The analyses looking at the effect of exit exams on graduation rates were not able to distinguish results for lower and higher performing students, though it is reasonable to expect that the requirements primarily affected lower performing students. Dee and Jacob (2007) also looked at college attendance, employment and earnings, and they found no overall effect from exit exam adoption.

EXPERIMENTS USING REWARDS

We identified causal studies related to 11 different experiments—in both the United States and in other countries—with rewards as the incentive for high performers. In the discussion below, we identify the experiments primarily grouped by location, in two cases clustering together several different but related experiments that were performed in the same location. The order of the discussion is alphabetical: India (one example), Israel (three examples), Kenya (two examples), Nashville (one example), New York City (one example), Ohio (one example), the Teacher Advancement Program in the United States (one example,), and Texas (one example).

Example 5: India

The Indian state of Andhra Pradesh conducted a 2-year experiment with teacher performance pay in rural elementary schools (Muralidharan and Sundararaman, 2011). The program randomly assigned schools to receive schoolwide incentives, individual teacher incentives, or to serve as a control group. The study also included two conditions that involved supplying extra resources in the form of either an additional teacher or cash for school materials. Each of the five conditions included 100 schools, with a typical school having three teachers and around 100 students. The performance pay in the two incentive conditions was based on average gains in student test scores in mathematics and language, measured either for the school as a whole in the schoolwide incentives condition or for the teacher's own students in the individual teacher incentives condition. The experiment used specially designed tests that explicitly included both basic and higher order skills,[20] and also included tests on science and social studies that did not receive incentives. The bonuses averaged about 3 percent of annual pay. The two incentive conditions did not include other types of support.

Averaged over the 2 years of the program, the test scores for the schools in the two incentive conditions were 0.19 standard deviations higher than the control schools (Muralidharan and Sundararaman, 2011, Table 3).[21] The effects in both years were positive and statistically significant. Scores increased in both subjects, though the difference was larger for mathematics than language. Scores were higher in the two incentive conditions for both lower and higher performing students, with no statistically significant interaction of the incentive effect with student baseline score (Table 6 and Figure 3).

[20] As a result of the use of both a broad test and an indicator based on gains (rather than a single proficiency level), this study was one of the few that has a "broad" performance measure within subjects (see Table 1).

The study did not include results on low-stakes tests in mathematics and language. However, scores on low-stakes tests in two subjects that were not a focus of the incentive program—science and social studies—were higher in the two incentive conditions, by an average of 0.14 standard deviations over 4 combinations of subject and year, with all 4 effects positive and statistically significant (Table 7). There was no difference between the effect of schoolwide and individual teacher incentives in the first year but the individual incentive schools performed statistically significantly higher in the second year. Over the 2 years, the average effect was 0.22 standard deviations for the individual incentives and 0.15 standard deviations for the schoolwide incentives (Table 8).[22]

The study included information about changes in teacher behavior that was obtained from direct observation and teacher interviews. Direct observation was conducted at each school several times during the 2 school years. There were no significant differences between the incentive and control schools in the direct observations measures of classroom process and teacher activity. In particular, the high level of teacher absenteeism—roughly 25 percent—was not affected by the incentives. In interviews, however, teachers in the incentive conditions reported higher levels of homework, classwork, instructional time, test preparation, and attention to lower performing students than did teachers in the control schools (Muralidharan and Sundararaman, 2009, Table 9). These reported differences were large and statistically significant in all cases, and in three cases were significantly correlated with student test scores.

The two resource conditions increased test scores by an average of 0.09 standard deviations over the 2 years of the program (Muralidharan and Sundararaman, 2009, Table 10).[23] The effect of the resource conditions over the control was statistically significant in both years, but the improvement in the resource conditions was also significantly lower than the improvement in the incentive conditions. The spending in the resource conditions was chosen to roughly equal the spending in the incentive conditions, so the higher increases in the incentive conditions suggests that they might have been more cost effective. However, it is likely that the test scores in the incentive conditions were inflated by the attachment of the incentives while the test scores in the resource conditions were not; as a result, a valid comparison of the incentive and resource conditions cannot be made.

Examples 6, 7, and 8: Israel

Three different experiments in Israel were conducted to provide incentives to increase the number of students passing the *bagrut*, a high school certification typically earned by students who intend to go to college (Angrist and Lavy, 2009; Lavy 2002, 2009). (The *bagrut* is comparable to college-bound certificates in other countries such as the baccalaureate in France and the A-levels in the United Kingdom.) Unlike most of the other incentive programs that we discuss, the tests that formed the basis for the experiments in Israel were voluntary and also involved some choice about subjects and levels of difficulty. As a result, the programs could potentially have affected the number and difficulty of the tests taken, as well as the passing rate. Students must receive a total of 20 credits to earn the *bagrut* certificate, with each test worth 1 to 5 credits, depending on its difficulty.

The first program—Example 6—provided schoolwide incentives to teachers in comprehensive high schools, a school that includes grades 7-12 and covers two-thirds of the Israeli population (Lavy, 2002). The rewards were distributed to all teachers in winning schools in proportion to their salaries, with the resulting

[21] We averaged the effects in the “Incentive School” row of Panel A for the columns that included school and household controls (Muralidharan and Sundararaman, 2011, Table 3). Our average included the results for the first and second year; the effects were 0.17 and 0.22 standard deviations, respectively.

[22] We averaged the effects in the “Individual Incentive School” and “Group Incentive School” rows, respectively, for column [1] for “Year 1 on Year 0” and for column [4] for “Year 2 on Year 0” in Table 8 of Muralidharan and Sundararaman (2011).

bonuses ranging from $250 to $1,000 at a time when the mean teacher salary was about $30,000. The program was designed as a tournament so that only schools in the top one-third received bonuses. Performance was based on three measures—credits earned per student, the proportion of students receiving the certificate, and the dropout rate—and was adjusted for the level of performance expected given the background of the students in the school. The 3-year program included 62 schools of the 170 comprehensive high schools in Israel. The typical school in the study had roughly 70-90 teachers and 500-1,500 students (Lavy, 2002, Table 1). The incentives program did not provide additional forms of support.[24]

The second program—Example 7—provided individual incentives to teachers in grades 10-12 who were teaching classes in English, mathematics, or another core subject that would prepare students for one of the *bagrut* tests (Lavy, 2009). Rewards were based on the passing rate and the mean score for each class, with an adjustment that reflected expected performance based on student and school characteristics. Teachers received bonuses if their classes exceeded the expected results by a large amount, with bonuses per class ranging from $1,750 to $7,500. Since some teachers prepared multiple classes for exams, the bonuses could be large relative to the mean teacher salary of $30,000. The program was implemented at 49 comprehensive high schools that typically had low numbers of students who received the *bagrut*. The program included 629 teachers: 302 teachers received rewards, 16 of whom received rewards for two classes. The high schools included in the program had a combined senior class size of roughly 7,000 students. The program was expected to last 3 years but was discontinued after 1 year because of budget cuts. The program also did not provide additional forms of support.

The third program—Example 8—provided monetary incentives to students for passing *bagrut* tests (Angrist and Lavy, 2009). The program was implemented in 20 nonvocational high schools with low proportions of students who receive a *bagrut*. The incentives included small rewards for continuing in high school to the eleventh and twelfth grades and for taking any of the *bagrut* tests. Larger rewards were given for passing the tests, with the largest reward given for earning the 20 credits needed for a certificate. Students who received all the awards would have received an amount equal to roughly four months of full-time work at a typical wage for high school dropouts and students who work during the summer. However, as with the teacher incentives program (Example 8), the student incentives program was planned for 3 years but discontinued after only 1 year, so no students were able to receive awards in multiple years. The program included about 4,000 students (Angrist and Lavy, 2009, Table 1). Like Examples 6 and 7, the program did not provide additional forms of support.

For Examples 6 and 7, the high schools were selected in a way that made it possible to define a set of untreated schools to use as a control group in order to be able to draw causal conclusions. For the program with schoolwide teacher incentives (Example 6), the proportion of students earning a certificate before the study was about 50 percent, and the program made no significant change in this overall proportion, though some specifications showed an increase of 3-4 percentage points, which approached significance (Lavy, 2002, Tables 1 and 2). Over 8 combinations of year, school type, and comparison group, the average increase was 2.2 percentage points.[25] None of the estimates of change in the

[23] We averaged the results in the “Inputs” row for “Year 1 on Year 0” in column [1] and “Year 2 on Year 0” in column [4] in Table 10 of Muralidharan and Sundararaman (2011).

[24] Lavy (2002) contrasted the effect of the school incentives program with the results of a program implemented in 22 high schools in which extra teachers were used to help improve performance on the *bagrut* tests.

certification rate was statistically significant; 6 were positive and 2 were negative. There were indications of increases in the proportion of students taking exams, the proportion achieving passing scores and the number of credits earned, though in the first year these increases appeared only for religious schools. Over 8 combinations of year, school type, and comparison group, the average effect of the incentives program on test scores was 0.11 standard deviations.[26] Six of the effect sizes were positive and statistically significant; two were not statistically significant, of which one was positive and one was negative. Information about contrasts in program effects for lower and higher performing students was not provided; however, using parents' schooling and family size as proxies for student performance, the analysis showed that the program effects were concentrated on lower performing students (Lavy, 2002, Table 3).

For the program that provided individual teacher incentives (Example 7), the analysis showed increases in both tests attempted and passed, as well as average scores (Lavy, 2009, Table 4). Over 2 subjects, math and English, the effect of the program on test scores averaged 0.19 standard deviations.[27] Using standard errors clustered by school and year, the effects were statistically significant for both subjects. Looking separately at students by quartile, there were positive and statistically significant effects on average test scores for the bottom three quartiles in math and the bottom quartile in English. For the top quartile in math and the top three quartiles in English, the effect on average test scores was small, mixed in sign, and not statistically significant (Table 4). The report on Example 7 did not provide information about changes in the proportion of students earning the *bagrut* certificate. A survey of teachers suggested that the incentives might have caused a number of changes in teaching methods and effort, including the use of individualized and small-group instruction, tracking students by ability, and the addition of instruction time, particularly before the tests (Table 8).

For the program that provided student incentives (Example 8), the 20 high schools selected for the program were chosen randomly from a pool of 40 low-performing schools so that the schools not chosen for the program were an equivalent comparison group. The analysis focused on students who were seniors in the single year that the program operated, since these were the only students for whom the program operated as planned. The program produced a statistically significant increase in the proportion of girls earning a *bagrut* certificate of 10 percentage points; there was no effect for boys (Angrist and Lavy, 2009, Table 2, Panel A).[28]

When the effects for girls and boys were pooled together, the average increase in earning a certificate was 5.4 percentage points over 8 different model specifications, with all effects positive but none statistically significant (Table 2, Panel A, columns 1 and 2). The effect for girls was concentrated in the higher performing students; in these low-performing schools, only 50 percent of the higher performing girls received a certificate without the program and the incentives increased the proportion for these girls by about 20 percentage points (Table 4, Panel A, column 3). As with the other two programs, the student incentives increased both credits attempted and credits earned (Table 7, column 7). Surveys of students showed no effect of the program on study time, study effort, or paid employment, but the higher performing girls—the group for whom the program effect was concentrated—did show a statistically significant increase in participation in the marathon study sessions that are commonly held around the spring holidays (pp. 1,403-1,404).

Examples 6 and 8 point to a consistent finding that incentives can be used to increase the proportion of students earning a *bagrut* certificate, with the effect concentrated among students who are on the borderline of receiving a certificate. The effect was stronger in Example 8, using student incentives, which was tar-

[25] We averaged the figures in columns [9] and [10] of Table 2 in Lavy (2002).

[26] We converted the score effects in columns [5] and [6] of Table 2 with the test score standard deviations of 21.088 and 19.780 for religious and secular schools, respectively, reported in Table 1 (Lavy, 2002).

[27] We converted the estimates in the "Treatment effect" row of the "Average score" section of Table 4 of Lavy (2009) into effect sizes by dividing by the average of the two test score standard deviations reported in the previous footnote from Lavy (2002). We used the estimates in columns (2) and (8) that cover all quartiles and use full controls.

geted for high schools with low proportions of students earning a certificate: in that setting the affected students were in groups where 40-50 percent of the students earned a certificate. With the program with schoolwide incentives (Example 6), there was a weaker response, probably because the program included a wide range of schools: some were far below the 40-50 percent level where few students have a realistic chance of earning a certificate; and others were far above the 40-50 percent level, where most students would be expected to earn a certificate.[29] There was evidence that the incentive programs produced changes in the behavior of teachers and students, with more focused instruction by teachers and increased effort by both teachers and students, primarily related to test preparation. The three programs did not include any low-stakes tests in the tested subjects to see whether the increased performance on the *bagrut* tests corresponded to more generalized achievement in those subjects.

Examples 9 and 10: Kenya

Two different experiments were conducted in primary schools in rural Kenya, one using schoolwide incentives to teachers and the other using incentives to students and their parents (Glewwe et al., 2010; Kremer et al., 2009). Both programs operated for 2 years. Primary schools in Kenya go through eighth grade, with a national test at the end that determines whether or not students go on to secondary school. Dropout rates in grades 5-7 are generally high in the country, with girls dropping out at a higher rate, and only one-third of all students finish the eighth grade (Kremer et al., 2009, p. 438).

The first incentives program—Example 9—provided schoolwide incentives to teachers on the basis of the students' average performance on district tests in grades 4-8 in seven subjects (Glewwe et al., 2010). Payments were given to schools that achieved either high scores or high score gains in a tournament across all the schools in the program. Although the performance indicator used gains, we coded the performance measure within subjects as "narrow" in Table 1 because the district tests relied solely on multiple choice questions (Glewwe et al., 2010, p. 211). The program included 50 schools, and 24 schools received prizes. In the winning schools, teachers of students in the tested grades received equal prizes, according to the school's rank in the tournament, with the prizes ranging from 21 to 43 percent of teachers' average monthly salary. The typical school had 12 teachers and 200 students, with roughly half in the grades affected by the program. No additional support was given as part of the program.

The second incentives program—Example 10—provided awards to students and their parents on the basis of students' performance on district tests in grade 6 in five subjects (Kremer et al., 2009). The program focused on girls, with the goal of increasing primary school completion among higher achieving girls. Prizes were given to the top 15 percent of girls according to their overall scores on the district exams. Winners were given money to pay for school fees and supplies for seventh and eighth grade. The program included 64 schools chosen randomly from a larger set (3 schools withdrew during the first year). Of the treated schools, 36 had at least one winner in the first year of the program, and 43 had at least one winner in the second year. No additional support was given as part of the program.

Examples 9 and 10 both randomly selected participating schools from a set of eligible schools so there was

[28] Angrist and Lavy (2009) referenced a number of studies on financial incentives in education that show stronger responses of females than males.

[29] In the program with schoolwide incentives, the standard deviation across schools in the proportion of students earning a certificate was about 50 percentage points (Lavy, 2002, Table 1).

an experimental comparison group for analysis. In the program with schoolwide incentives to teachers (Example 9), test scores on the district exams were not significantly different during the first year; however, in the second year, the test scores increased by 0.14 standard deviations more than the comparison schools (Glewwe et al., 2010, Table 3, Panel A, columns 5 and 6). Over the 2 years, the average effect size on the high-stakes tests was 0.09 standard deviations. Low-stakes tests given by the organization sponsoring the experiment were mixed in sign and showed no significant effects, with an average effect of 0.01 standard deviations (Table 3, Panel B, columns 5 and 6). The district tests used in the program did not show any statistically significant increase in scores in program schools the year after the incentives ended, though the effect was positive (Table 3, Panel A, column 7).

Consistent with the incentives, which assigned a low score to students who did not take the exam, the first program increased the number of students taking the district tests by 7 percentage points averaged over the 2 years (Glewwe et al., 2010, Table 2, Panel B, columns 2 and 3). The program did not result in any significant changes in teacher attendance, homework assignment, or various measures of instruction (which were coded by trained observers) (Table 5, columns 4 and 5).[30]

In the program with incentives to students and their parents (Example 10), test scores on the district tests for girls increased by 0.12-0.19 standard deviations (Kremer et al., 2009, Table 4). The implementation of the program in one of the two districts was marred by low levels of trust with the sponsoring organization and a fatal lightning strike in a primary school; an analysis restricted to the Busia district, which did not experience these problems, showed an increase of 0.19-0.27 standard deviations in the district tests. Over 6 combinations of district, baseline control, and sample, the average effect size on the high-stakes tests was 0.20 standard deviations, with 4 of the effects statistically significant.[31] The test score effects occurred for both lower and higher performing girls within the schools (Kremer et al., 2009, p. 447). In Busia, the increased performance by the first cohort of girls on the district tests in the year of the program continued in their performance on the district tests the following year, when they were no longer in the program (0.24 standard deviations, p. 452), with the program affecting both girls who won prizes and girls who did not. The Busia girls in the first cohort also took a low-stakes test given by the sponsoring organization in the year after being in the program, and they showed an increased performance of 0.19 standard deviations above the girls who had been in control schools (p. 452). Both of these effects in the year following participation in the program by the first cohort Busia girls were statistically significant. A survey about attitudes related to education found no evidence that the incentives program affected student motivation (Table 8, Panel A).

There was some suggestion that the second program also improved outcomes for boys as well, even though they were not the focus of the program (Kremer et al., 2009, Table 5). There was no indication of significant program impacts on student attitudes, study habits, or available educational materials (Table 8). Unlike the school incentive programs in Kenya and the school and teacher incentive programs in India, the student and parent incentives program in Kenya increased teacher attendance by about 5 percentage points (Table 7).

Example 11: Nashville

[30] As with the schools in the incentive programs in India, the teachers in the programs in Kenya had a high rate of absenteeism, averaging roughly 20 percent (Glewwe et al., 2010, p. 206).

[31] We averaged the program school estimates for the Busia and Teso districts combined and the Busia district separately, for analyses for the intention to treat (ITT), restricted, and longitudinal samples of Table 4 of Kremer et al. (2009). For the restricted sample estimates, we used the analysis with controls for mean school test scores in the year before the program began. For the longitudinal sample estimates, we used the analysis with controls for individual school test scores in the year before the program began.

A 3-year experiment conducted in the Metropolitan Nashville School System provided incentive pay to middle school mathematics teachers (Springer et al., 2010). A total of 296 teachers volunteered to participate in the experiment and were randomly assigned to treatment and control groups. Teachers in the treatment group were eligible to receive annual bonuses of $5,000-$15,000 on the basis of a value-added measure of change in the test scores of their students on the Tennessee state mathematics test. Although the performance indicator used changes in test scores rather than a single proficiency target, we coded the performance measure within subjects as “narrow” because the Tennessee state tests used only multiple-choice questions for mathematics.[32] The performance levels for receiving a bonus were set between the 85th and 95th percentiles of the districtwide distribution for the value-added measure. The proportion of participating teachers who received a bonus increased from one-third in the first year to one-half in the third year (Springer et al., 2010, Table 1). Over the course of the experiment, half of the teachers became ineligible to continue participating in the program, in most cases because they stopped teaching middle school mathematics in the district (Table 3). No additional support was provided as part of the incentives program.

Over 3 years and four grades, the average effect of the incentive program was 0.04 standard deviations on the high-stakes test, which was not statistically significant (Springer et al., 2010, p. 29). Over all 12 combinations of year and grade, the effects were positive in 7 of 12 cases, and 2 of them were statistically significant; of the 5 cases with negative effects, none of them was statistically significant (Table 7). For grades 5 and 6 the effects were all positive; for grades 7 and 8 all effects but one were negative. The effect for grade 5 was statistically significant in two of three cases. The students in grade 5 in the second year of the experiment, associated with one of the two significant effects in grade 5, did not perform significantly differently in mathematics the following year (p.30). The study also looked at effects in reading, science, and social studies for the students of teachers in the experiment. There were no statistically significant effects for reading, but there were some statistically significant effects for science and social studies in grade 5, the same grade for which statistically significant effects appeared for mathematics (Springer et al., 2010, Tables C-1 to C-3).

Example 12: New York City

Fryer (2010) reports results of student incentive experiments carried out over 2 years in four urban school districts. In one of the districts—New York City—students were provided incentives on the basis of 10 interim tests in reading and mathematics that were designed to provide information related to the state standards and exams.[33] The fourth graders in the study could earn up to $25 on each test, and the seventh graders could earn up to $50 on each test, with the reward based on the score. The average fourth grader earned $139.43 and the average seventh grader earned $231.55 (Fryer, 2010, Table 1). Because the tests were designed to mirror the state exams, which include extended response items,[34] we coded the performance indicator as “broad.” A total of 63 schools were randomly chosen to participate in the experiment out of 143 volunteer schools that included more than 17,000 students.

[32] See Hightower (2010), state table for Standards, Assessments, and Accountability.

[33] In the other cities, the incentives were based on grades (Chicago), books read (Dallas), or attendance and behavior (Washington, DC). In Chicago, the effect of incentives based on grades was negative but small and not statistically significant (Fryer, 2010, Table 2). In Dallas, the effect of incentives based on books read was large and statistically significant for English speakers for measures of reading comprehension and language use but not vocabulary (Table 3). In Washington, DC, the effect of incentives based on attendance and behavior was moderate and positive but of marginal statistical significance (Table 3).

[34] See Hightower (2010), state table for Standards, Assessments, and Accountability.

The study reports the effect of the incentive program on the New York state tests in reading and mathematics.[35] Over eight combinations of subject, grade, and specification, the average effect size for the incentive programs was 0.01, with the effect sizes evenly distributed between positive and negative effects; none was statistically significant.[36] Considering the effects separately by subject, the average effect size was 0.00 for reading and 0.03 for mathematics, with each subject having two positive and two negative effects. Considering the effects separately by grade, the average effect size was 0.03 for fourth grade and 0.00 for seventh grade, with each grade having two positive and two negative effects. A separate assessment of student interest and enjoyment in schoolwork did not find a statistically significant change in motivation from the program, but the measured change was negative (Table 7).

Example 13: Ohio

A 3-year experiment in Coshocton, Ohio, a disadvantaged community, paid elementary school students in grades 3-6 for their scores on the state accountability tests in five core subjects (Bettinger, 2010). Students were paid $15 for each score at or above the 75th percentile and $20 for each score at or above the 85th percentile. All of the four elementary schools in Coshocton participated in the program at some time. The schools included roughly 900 students. No additional supports were provided by the program.

With four participating grades and four elementary schools, there were 16 grade-school combinations, half of which were randomly chosen each year to receive incentives under the program. The program resulted in a statistically significant increase of 0.13-0.19 standard deviations in the scores of the mathematics tests attached to the incentives (Bettinger, 2010, Table 3), but the effects on scores in reading, science, and social studies were small and not statistically significant, though all but one were positive (Tables 6 and 7). Information was not provided on the effect of the program on the writing test. Over 14 combinations of subject and model specification, the average effect on the high-stakes test was 0.06, with 4 of the 14 effects positive and statistically significant.[37] The effect in mathematics was concentrated on the lowest and highest quartile (Table 5).

[35] Given our criteria for coding the tests, we coded this as an example of a "low-stakes" test, since the state tests were not the tests that were being attached to the incentives in the experiment.

[36] We used the New York City estimates in Table 2 on the lines "Reading: All Controls" and "Math: All Controls" of Fryer (2010).

[37] We averaged the coefficients in the "Treatment" line of Tables 3, 6, and 7 of Bettinger (2010).

The study did not provide results for a low-stakes test. The study checked for spillover effects on siblings of students in classrooms with incentives: over four combinations of subject and model specifications, the effects on the siblings were consistently negative but none approached statistical significance (Table 10). Measures of changes in student motivation for academic tasks found no significant effects (p. 16).

Example 14: Teacher Advancement Program

The Teacher Advancement Program (TAP) is a comprehensive school reform model for the United States, developed by a foundation, that includes teacher performance pay (Glazerman et al., 2009; Glazerman and Seifullah, 2010; Springer et al., 2008). The performance award is based on value-added measures of the test score gains on the state achievement tests in both the teacher's individual class and averaged across the entire school, in addition to classroom observations by certified evaluators. Because the performance indicator includes both test score gains and classroom observations, we coded the performance measure within subjects as "broad." Rewards per teacher range up to $12,000, though the exact structure of the program varies by location (Springer et al., 2008).[38] As of 2007, the program had been implemented in more than 180 schools across the country, which includes roughly 5,000 teachers and 60,000 students. In addition to performance pay, TAP includes professional development and a system of multiple career paths to allow teachers to take on mentoring roles.

Example 14A: TAP in Chicago

Glazerman and colleagues studied the implementation of TAP in Chicago—Example 14A—using a hybrid experimental design in which treated schools were randomly assigned to year of implementation and were also matched to non-TAP control schools (Glazerman et al., 2009; Glazerman and Seifullah, 2010). Thus far, there are results for 2 years for the first cohort of schools and 1 year for the second cohort of schools. There were eight TAP elementary (K-8) schools in each cohort.[39] The studies analyzed changes in the test scores of the tests attached to the incentives. The first-year study found effect sizes of −0.04 for both reading and mathematics, but neither effect was statistically significant. Across the 10 combinations of subject and grade in the study, 2 of the 10 effect sizes were positive and 8 were negative, and none was statistically significant (Glazerman et al., 2009, Tables IV.1 and IV.2). The second-year study found effect sizes of 0.00 for reading and 0.02 for mathematics, neither of which was statistically significant. Across the 10 combinations of subject and grade, 6 of the effect sizes were positive and 4 were negative, with none being statistically significant. (Glazerman and Seifullah, 2010, Tables III.1 and III.2).

The studies also looked at the effect of TAP on teacher retention.[40] In the first year, the first cohort showed a statistically significant increase in teacher retention at the school level of 5.2 percentage points (Glazerman et al., 2009, Table IV.5); this increase was concentrated in academic teachers who were not in the tested grades and subjects. In the second year, there was an increase in retention of 1.0 percentage point, which was not statistically significant (Glazerman and Seifullah, 2010, Table IV.1).

Example 14B: A Comparison of Mathematics Test Scores

Another study of TAP (Springer et al., 2008)—Example 14B—compared mathematics test score growth in schools that implemented TAP and schools that did not, using two different ways of controlling statistically for unobservable differences between the two types of schools. Over a 4-year period, the study analyzed data in two states for 1,200 schools in which 28 schools implemented TAP. To measure achievement growth, the study used fall-to-spring gains on the Northwest Evaluation Association (NWEA) tests in mathematics, given in grades 2-10, which were not attached to the incentives program. In grades 2-5, TAP schools increased test score gains by 1-2 points (Springer et al., 2008, Tables 6-7). The gains were statistically significant and correspond to an effect of roughly 0.2 standard deviations on gains that typically have a standard deviation of 7-8 points. In grades 6-8, the changes in TAP schools were small and mixed, with the only statistically significant changes being decreases of about 1 point for two grades in one of the two models. In grades 9-10, both models showed statistically significant decreases of 1-3 points. Over 18 combinations of grade and model specification, the aver-

[38] In the Chicago implementation of TAP, performance pay was phased in so that it was smaller during the first year of the program than it was in the second year. In the first cohort of schools, the first year bonus averaged $1,100, ranging from $0 to $2,045, and the second year bonus averaged $2,653, ranging from $0 to $6,320 (Glazerman and Seifullah, 2010, Table I.1).

[39] The TAP implementation in Chicago also included two high schools in each year, but the studies did not analyze their results because of the difficulty in finding appropriate controls.

[40] The focus of the analysis appears to be on retention resulting from the effects of voluntary turnover, not retention resulting from involuntary personnel decisions.

age effect was 0.01 standard deviations, with 13 of the 18 effects statistically significant, 7 of them positive and 6 of them negative.[41]

Example 15: Texas

A nonprofit organization in Texas started a program in 1996 that provides rewards to students and teachers for scores on advanced placement (AP) course exams (Jackson, 2010). As of 2007, the program included more than 40 secondary schools with high numbers of disadvantaged students. AP teachers receive payments of $500-$1,000 for each of their students who earns a score of 3 or higher on the AP test. Students receive a bonus of $100-$500 for each score of 3 or higher. Students must be enrolled in the corresponding AP course in order to earn the bonus from an AP test. The program also provides bonuses to teachers for being part of the program, ranging from $500-$1,000 for teachers in pre-AP courses to $3,000-$10,000 for the lead teachers who organize and provide training for the AP program in a school. In addition to the financial rewards, the program includes teacher training, as well as a curriculum for the earlier grades to help prepare students for AP courses. Support for the program is provided primarily by private donors, who have some role in selecting a school and choosing which AP subjects will be rewarded and how large the rewards will be. The subjects typically included in the program are English, mathematics, and one or more of the sciences.

Jackson (2010) compared changes in outcomes in schools that adopted the AP incentive program to the changes in outcomes in schools that had chosen to adopt the program but had not yet done so because no donor had been found. The analysis measured student achievement with SAT and ACT test results, using a criterion of 1,100 on the SAT and 24 on the ACT. In schools selected for the program, 20 percent of graduates met the criterion on the SAT or ACT in the preferred model (Jackson, 2010, Table 2, model 28). In the schools that implemented the program, the proportion of graduates who met the criterion increased by 2 percentage points the first year and by 1 additional percentage point each in the second and in third years (Table 7, column 1). There was no significant change in the number of students who took the SAT or ACT (Table 2, model 22). There was no significant increase in AP course enrollment for the first 2 years of the program, but starting in the third year, enrollment increased by 34 percent (Table 3, column 1). There was an increase of 1.2 percent in the graduation rate, but the result was not statistically significant (Table 2, model 16). However, the number of students attending college increased by 5.3 percent (Table 2, model 34).

CONCLUSIONS

In this section we synthesize the results across the different incentive programs discussed above and summarized at the end of this chapter in Tables 4-1A, 4-1B, 4-2, and 4-3. We focus specifically on summarizing the types of incentive programs investigated and analyzing the effect of those programs on student achievement and on high school graduation and certification. We then consider the relative costs and benefits of incentive programs.

[41] We computed the average from the coefficients on the "TAP" line of Tables 6 and 7 in Springer et al. (2008) and then divided by a standard deviation of 7.5 because the NWEA tests in the elementary grades have a standard deviation of 7-8 points (p. 11). We did not have direct information about the standard deviation of the NWEA tests in the upper grades and so used 7.5 as the estimate for all grades. We did not use the results in Table 5, which did not control for selection of schools into the program and therefore did not support a causal interpretation about its effect.

Types of Incentive Programs Investigated in the Literature

As summarized in Tables 4-1A and 4-1B, researchers and policy makers have explored incentive programs with a relatively wide range of variation in key structural features. Across the 15 examples we analyzed, there are substantial differences in who receives incentives, the breadth of the performance measures across and within subjects that are attached to the incentives, the nature of the consequences that the program attaches to the performance level, and whether extra support is provided by the program. In addition, there are differences in the nature and frequency of the consequences attached to the performance measures that are summarized in the text describing the programs, though not coded in the table.

The research literature we reviewed (see Chapters 2 and 3) suggests that these key structural features could be critical to the successful operation of an incentive program, so it is notable that the literature includes examples of different options for the different features. Looking at the feature options one at a time, the studies we review provide examples of major contrasts that could potentially be important, and for each contrasting feature option in the table, there are at least several strong studies that investigate programs containing that option.

When we considered the feature options in combination, however, it is clear that many possible combinations of the basic structural features do not appear: see Tables 4-1A and 4-1B. Some unexplored combinations are likely to seem uninteresting to implement as actual programs—such as a possible incentive program that might combine consequences in the form of sanctions while providing no additional support, which would likely prove to be politically untenable. However, there are a number of unexplored feature combinations that are potentially interesting and seem potentially promising for implementation and study.

In the current policy context, there are at least two such unexplored combinations of structural features that are salient: the combination of incentives for schools and broad performance measures within subjects, and the combination of incentives for individual teachers and sanctions.

The first combination is a frequently mentioned possible change that might be introduced with the next reauthorization of the Elementary and Secondary Education Act (ESEA)—school accountability with performance measures that have broader coverage within subjects by using tests that better reflect higher order thinking skills and indicators that are sensitive to changes across a broader range of performance than a single proficiency level.

The second combination is a frequently mentioned possible change in discussions about teacher quality—incentives for individual teachers in the form of sanctions that require teachers whose students do not meet some test-based level of performance to leave the profession (see, e.g., Lang, 2010; Staiger and Rockoff, 2010). Proposals to use the results of student tests as an input into teacher tenure decisions—which can be interpreted as subjecting teachers to a strong sanction if their students perform poorly—are an example of this combination. We do not take a position on either of these proposals here or on other unexplored combinations that may be proposed. Instead, we note the twin points that the existing research literature contains information about the effects of incentive programs that use these features in other combinations, but it does not contain information about the effects of programs with these particular combinations of features.

Effects on Student Achievement and High School Graduation and Certification

We summarize the effects of the incentive programs on student achievement and high school graduation and certification in Tables 4-2 and 4-3. We discuss these effects in terms of four groupings of programs: NCLB and its predecessors, high school exit exams, programs using rewards in other countries, and programs using rewards in the United States.

NCLB and Its Predecessors

The four studies that we analyzed all provided information about the achievement effects of test-based incentives targeted at schools that are in the NCLB mold.[42] The studies showed average incentive effects on the low-stakes tests ranging from 0.04 to 0.22 standard deviations. Across the studies there were a number of individual effect estimates that were positive and statistically significant, though there were also many that were not statistically significant and some that were negative.

At first blush, the evidence of incentives on student achievement from these studies appears substantial. However, there are two important caveats. First, the statistically significant effects were concentrated in fourth grade math; in contrast, the results for eighth grade math and for reading for both grades were often not statistically significant and sometimes negative.

Second, the highest two estimates—0.22 and 0.12 standard deviations—were problematic. Both estimates came from analyses that excluded results for eighth grade reading, giving an unbalanced overall picture of the effects of the incentives on achievement. In addition, the highest estimate of 0.22 standard deviations came from comparisons between public and private schools that may have been affected by movement away from Catholic schools that occurred during the early years of NCLB. Without these two problematic estimates, the effects estimated by the research range from 0.04 to only 0.08 standard deviations.

Given these two caveats, the evidence related to the effects on achievement of test-based incentives to schools appears to be modest, limited in both size and applicability. Our preferred estimate for these programs is 0.08 standard deviations, reflecting the national results for both the pre-NCLB period by Lee (2008) and the NCLB period by Dee and Jacob (2011). A program with an effect size of 0.08 standard deviations would raise the achievement of students currently at the 50th percentile to the 53rd percentile. This gain is small, both by itself and in comparisons across nations: the highest achieving countries on international tests often perform a full standard deviation above the United States, measured in terms of the distribution of performance within the United States (see, e.g., Gonzales et al., 2008, Figure 14 for TIMSS 2007 mathematics). To achieve an increase of the magnitude needed to match the high performing countries would mean that students currently at the 50th percentile in the United States would have to increase their scores to the current 84th percentile. For underachieving groups, far more improvement would be needed because of the large achievement gaps in the United States (Hill et al., 2008, Table 2). Although an effect size of 0.08 standard deviations is small in comparison with the improvements the nation hopes to achieve, it is comparable to the effect sizes found for other promising interventions that have been evaluated using standardized tests with relatively broad subject coverage (Hill et al., 2008, Table 4). The influential Tennessee STAR experiment with class-size reduction was notable for achieving effect sizes ranging from 0.15 to 0.25 standard deviations (Finn and Achilles, 1999), though the gains from class-size reduction have been much smaller when they were instituted on a statewide basis (e.g., Stecher et al. 2001).

High School Exit Exams

[42] One of the research papers was a meta-analysis covering 14 studies, many of which would meet our inclusion criteria if we had considered them separately.

One of the three studies on the effects of high school exit exam requirements provided estimates of the effects on achievement on a low-stakes test: it found an average effect of 0.00 standard deviations (see Table 4-2). The other two studies provided estimates of the effects on graduation: they found average effects of −2.1 and −0.6 percentage points (see Table 4-3). A number of the negative effects are statistically significant. The smaller estimate was for a study that counted GEDs as equivalent to high school diplomas; excluding this study leaves an estimate of the graduation effect of −2.1 percentage points.

Incentive Programs That Use Rewards in Other Countries

The committee's analysis included six studies of incentive programs that used rewards in other countries, in India, Israel, and Kenya. The Kenya study measured the effect of incentives on achievement using low-stakes tests, while the studies in India and Israel measured the achievement effect using the tests attached to the incentives (see Table 4-2). The six studies found average estimates of the effect on achievement ranging from 0.01 to 0.19 standard deviations, and most of the high positive effects are statistically significant. Two of the Israel studies found effects on high school certification that averaged 2.2 and 5.4 percentage points (see Table 4-3). The Israel studies found that the effects on both achievement and certification were concentrated on lower-performing students.

As with the studies on NCLB and its predecessors, the studies on foreign reward programs suggest substantial benefits of incentive programs that must be considered in light of important caveats. First, the programs in India and Israel measured achievement using the high-stakes tests attached to the incentives. The problems with this measure are discussed above, and it is not clear how much change in achievement would be shown on low-stakes tests.

Second, the programs in India and Kenya were in developing countries that have quite a different context for education than that in developed countries. In particular, the high level of teacher absenteeism and the high rate of student dropout in middle school suggest that the incentives for both teachers and students may operate differently in developing countries.

Given these caveats, it is not clear what can be learned from these studies that would be applicable to the use of incentives in the United States. For all three countries, there are difficulties in drawing conclusions about the ability of such programs to increase achievement in the United States. In addition, although the ability of the Israel programs to increase high school certification with incentives is potentially promising, it is hard to evaluate the value of the increase without knowing whether it is accompanied by increased learning beyond that measured by the high-stakes test.

U.S. Incentive Programs That Use Rewards

Six of the seven studies that provided information about U.S. incentive programs that use rewards showed average effects on achievement that ranged from −0.02 to 0.06 standard deviations (see Table 4-2). Many effects were positive, and some were statistically significant, but there were also a number of negative effects. The estimates of achievement effects included a number that were based on the tests attached to the incentives; when these are eliminated, there are two studies, both of which found 0.01 standard deviations. One study showed an effect of incentives on high school graduation of 0.9 percentage points, but the effect was not statistically significant (see Table 4-3).

On the basis of our synthesis of the evidence, summarized above, we reached two conclusions about the effect of test-based incentives on student achievement and high school completion.

> **Conclusion 1: Test-based incentive programs, as designed and implemented in the programs that have been carefully studied, have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests, which are less likely to be**

inflated by the incentives themselves, the overall effects on achievement tend to be small and are effectively zero for a number of programs. Even when evaluated using the tests attached to the incentives, a number of programs show only small effects. Programs in foreign countries that show larger effects are not clearly applicable in the U.S. context. School-level incentives like those of the No Child Left Behind Act produce some of the larger estimates of achievement effects, with effect sizes around 0.08 standard deviations, but the measured effects to date tend to be concentrated in elementary grade mathematics and the effects are small compared to the improvements the nation hopes to achieve.

Conclusion 2: The evidence we have reviewed suggests that high school exit exam programs, as currently implemented in the United States, decrease the rate of high school graduation without increasing achievement. The best available estimate suggests a decrease of 2 percentage points when averaged over the population. In contrast, several experiments with providing incentives for graduation in the form of rewards, while keeping graduation standards constant, suggest that such incentives might be used to increase high school completion.

Balancing the Benefits and Costs of Test-Based Incentives

The research to date suggests that the benefits of test-based incentive programs over the past two decades have been quite small. Although the available evidence is limited, it is not insignificant. The incentive programs that have been tried have involved a number of different incentive designs and substantial numbers of schools, teachers, and students. We focused on studies that allowed us to draw conclusions about the causal effects of incentive programs and found a significant body of evidence that was carefully constructed. Unfortunately, the guidance offered by this body of evidence is not encouraging about the ability of incentive programs to reliably produce meaningful increases in student achievement—except in mathematics for elementary school students.

Although the evidence to date about the effectiveness of incentive programs has not been encouraging, the basic research findings suggest a number of features that are likely to be important to the effectiveness of incentive programs and that can provide guidance in the design of new models. Some proposals for new models of incentive programs involve combinations of features that have not yet been tried to a significant degree, such as school-based incentives using broader performance measures and teacher incentives using sanctions related to tenure. Other proposals involve more sophisticated versions of the basic features we have described, such as the "trigger" systems discussed in Chapter 3 that use the more narrow information from tests to start an intensive school evaluation that considers a much broader range of information and then provides more focused supports to aid in school improvement.

It is also likely to be important to consider potential programs that focus more on the informational role that tests can play. Our study has specifically *not* focused on policies and programs that rely solely on information about educational achievement that tests provide to drive improvement through educator motivation and public pressure. Our focus for the study was chosen because so much of the educational policy discussion over the past decade has been driven by the conclusion that mere information without explicit consequences is insufficient to drive change. And yet the guidance coming from the basic research in psychology suggests that the purely informational uses of test results may be more effective in some situations than incentives that attach explicit consequences to those results. As policy makers and educators continue to look for successful routes to improving education in the years ahead, the exploration should include more subtle incentives that rely on the informational role of test results and broader types of accountability.

In continuing to explore promising routes to using test-based incentives, however, policy makers and educators should take into account the costs of doing so. Over the past two decades, the education policy and research communities have invested substantial attention and resources in exploring the use of test-based incentives as a way to improve education. This investment seemed to be worthwhile because it appeared to offer a promising route for improvement. Further investment in test-based incentives still seems to be worthwhile because there are now more sophisticated proposals for using test-based incentives that offer hope for improvement and deserve to be tried. However, in choosing how much attention and investment to devote to the exploration of new forms of test-based incentives, it is important to remember that there are other aspects of improving education that also would benefit from development. In addition to test-based incentives, investments to improve standards, curriculum, instructional methods, and educator capacity are all likely to be necessary for improving educational outcomes. Although these other aspects of the system are likely to be complements to test-based incentives in improving education, they are competitors for funding and policy attention. Further research and development of promising new approaches to test-based incentives need to be balanced against the research and development needs of promising new approaches in other areas related to improving education. We have not considered those tradeoffs in our examination of test-based incentives, but those trade-offs are the most important costs that need to be considered by the policy makers who will decide which new incentive programs to support.

TABLE 4-1A Overview of Results from All Studies of Test-Based Incentive Programs Using Causal Analyses

		Structure of Incentives System[a]				
Incentive Programs		Target Who Receives Incentives	Perf Measure Across Subjects	Perf Measure Within Subjects	Consequences	Support
Studies of NCLB and Its Predecessors						
1.	U.S. pre-NCLB	Schools	Mixed	Mixed	Mixed	Mixed
2A.	U.S. NCLB	Schools	Narrow	Narrow	Sanction	Yes
2B.	U.S. NCLB	Schools	Narrow	Narrow	Sanction	Yes
2C.	U.S. NCLB	Schools	Narrow	Narrow	Sanction	Yes
3.	Chicago pre-NCLB	Schools and Students	Narrow	Narrow	Sanction	Yes
Studies of High School Exit Exams						
4.	U.S. HS Exit	Students	Mixed	Narrow	Sanction	Yes
Studies of Incentive Experiments Using Rewards						
5.	India	Teachers-I or Teachers-G	Narrow	Broad	Reward	No
6.	Israel Teachers-G	Teachers-G	Broad	Narrow	Reward	No
7.	Israel Teachers-I	Teachers-I	Broad	Narrow	Reward	No
8.	Israel Student	Students	Broad	Narrow	Reward	No
9.	Kenya Teachers-G	Teachers-G	Broad	Narrow	Reward	No
10.	Kenya Student	Students and Parents	Broad	Narrow	Reward	No
11.	Nashville	Teachers-I	Narrow	Narrow	Reward	No
12.	New York	Students	Narrow	Broad	Reward	No
13.	Ohio Student	Students	Broad	Narrow	Reward	No
14A.	TAP-Chicago	Teachers-I and Teachers-G	Broad	Broad	Reward	Yes
14B.	TAP-2 states	Teachers-I and Teachers-G	Broad	Broad	Reward	Yes
15.	Texas AP	Teachers-I and Students	Narrow	Narrow	Reward	Yes

NOTE: Teachers-G = Teachers-Group, Teachers-I = Teachers-Individually.

[a] The features related to the structure of incentive programs that should be considered when designing the programs are (1) the target for the incentives (schools, teachers, or students in these examples); (2) the extent to which the performance measures are aligned with the outcomes desired (broad or narrow), both across and within subjects; (3) the consequences that the incentives provide (reward or sanction); (4) the support provided to reach the performance goals; and (5) the way the incentives are framed and communicated. The last feature is not included in the table because no studies consider it.

TABLE 4-1B Overview of Results from All Studies of Test-Based Incentive Programs Using Causal Analyses

Outcomes[a]					
Effect on High-	Effect on Low-	Effect on Other Subject	Effect on HS Grad or	Effect on Lower Perf	Effect on Higher

Incentive Programs	Stakes Tests	Stakes Tests	Tests	Cert	Students	Perf Students
Studies of NCLB and Its Predecessors						
1. U.S. pre-NCLB		+				
2A. U.S. NCLB		0/+	0		+/0	+/0
2B. U.S. NCLB		0/+				
2C. U.S. NCLB		0/+				
3. Chicago pre-NCLB	+	0/+/−	+		+	+/0
Studies of High School Exit Exams						
4. U.S. HS Exit		0		−/0	test 0	test 0
Studies of Incentive Experiments Using Rewards						
5. India	+		+		+	+
6. Israel Teachers-G	+			+/0	+	0
7. Israel Teachers-I	+				+	0
8. Israel Student				+	+	0
9. Kenya Teachers-G	+/0	0				
10. Kenya Student	+	+			+	+
11. Nashville	0/+		0/+			
12. New York	0					
13. Ohio Student	+/0				+/0	+/0
14A. TAP-Chicago	0					
14B. TAP-2 states		+/−/0				
15. Texas AP		+		0		+

NOTE: Teachers-G = Teachers-Group, Teachers-I = Teachers-Individually.

[a]*Results* of studies are characterized here as positive (+), negative (−), or not statistically significantly different from zero (0). The most lenient level of significance provided in the study is used, generally $p < 0.10$ or $p < 0.05$.

TABLE 4-2 Summary of Average Effects of Incentive Programs on Student Achievement Tests

	Test Outcome		Distribution of Test Outcome Effects Across Analyses			
Incentive Programs	Type of Stakes	Overall Effect Size[a]	+Sig	+Nonsig	−Nonsig	−Sig
Studies of NCLB and Its Predecessors						
1. U.S. pre-NCLB	Low	0.08		87%	11%	
2A. U.S. NCLB	Low	0.08	25%	50%	25%	0%
2B. U.S. NCLB	Low	0.12[b]	33%	67%	0%	0%
2C. U.S. NCLB	Low	0.22[c]	17%	83%	0%	0%
3. Chicago pre-NCLB	Low	0.04	83%	22%	22%	22%
Studies of High School Exit Exams						
4A. U.S. HS Exit	Low	0.00	0%	50%	50%	0%
Studies of Incentive Experiments Using Rewards						
5. India	High	0.19	100%	0%	0%	0%
6. Israel Teachers-G	High	0.11	75%	13%	13%	0%
7. Israel Teachers-I	High	0.19	100%	0%	0%	0%
9. Kenya Teachers-G	Low	0.01	0%	50%	50%	0%
10. Kenya Student	Low	0.19	100%	0%	0%	0%
11. Nashville	High	0.04	17%	42%	42%	0%
12. New York	Low	0.01	0%	50%	50%	0%
13. Ohio Student	High	0.06	29%	64%	7%	0%
14A. TAP-Chicago	High	−0.02	0%	50%	50%	0%
14B. TAP-2 states	Low	0.01	39%	11%	17%	33%

NOTE: Teachers-G = Teachers-Group, Teachers-I = Teachers-Individually.

[a]Effect size is presented in standard deviation units.

[b]Omits eighth grade reading.

[c]Omits eighth grade reading; uses comparison to private schools during period of fluctuating enrollment.

TABLE 4-3 Average Effects of Test-Based Incentive Programs on High School Graduation/Certification Rates

	Distribution of Rate Changes Across Analyses				
Incentive Programs	HS Grad/ Cert Rate Changes	+Sig	+Nonsig	−Nonsig	−Sig
Studies of High School Exit Exams					
4B. U.S. HS Exit	−2.1%	0%	0%	0%	100%

4C. U.S. HS Exit	−0.6%	0%	0%	33%	67%
Studies of Incentive Experiments Using Rewards					
6. Israel Teachers-G	2.2%	0%	75%	25%	0%
8. Israel Student	5.4%	0%	100%	0%	0%
15. Texas AP	0.9%	0%	50%	50%	0%

NOTE: Teachers-G = Teachers-Group.

The National Academies of SCIENCES ENGINEERING MEDICINE

The National Academies of Sciences, Engineering, and Medicine
500 Fifth St., NW | Washington, DC 20001