Monetary Incentives for Low-Stakes Tests

CSE Report 625

Harold F. O'Neil
University of Southern California / National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Jamal Abedi, Charlotte Lee, Judy Miyoshi, and Ann Mastergeorge
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

April 2004

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Motivational Incentives for Low-Stakes Tests

Project Directors: Harold F. O'Neil, University of Southern California / CRESST, and Jamal Abedi, CRESST / UCLA

The work reported herein was supported in part by the U.S. Department of Education under the American Institutes for Research (AIR) / Education Statistical Services Institute (ESSI) Contract Number RN95127001, Task Order 1.2.93.1, as administered by the National Center for Education Statistics (NCES), U.S. Department of Education. The work reported herein was also supported under the Educational Research and Development Centers Program, PR / Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, the U.S. Department of Education, the American Institutes for Research / Education Statistical Services Institute, or the National Center for Education Statistics.

MONETARY INCENTIVES FOR LOW-STAKES TESTS[1]

Harold F. O'Neil

University of Southern California/National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Jamal Abedi, Charlotte Lee, Judy Miyoshi, and Ann Mastergeorge[2]

University of California, Los Angeles/National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Abstract

Recent information on international assessments (e.g., the Third International Mathematics and Science Study) indicates that 12th-grade students in the United States are doing extremely poorly on such assessments compared with their peers in other countries. These poor results are usually attributed to cognitive factors such as students' opportunities to learn. However, a partial explanation of these results may be motivational. Because the low-stakes tests were administered in these 12th-graders' final year in high school, this timing may have negatively affected motivation, and thus performance. Using money as an incentive ($10.00 per item correct), on a test using TIMSS released math items, we manipulated the amount of money per item correct so as to increase a motivational effect and thus increase performance. A focus group, pilot study, and main study were conducted. The monetary incentive was not effective in improving performance.

Recent information on international assessments (e.g., the Third International Mathematics and Science Study [TIMSS]) indicates that 12th-grade students in the United States are doing extremely poorly on such assessments compared with their peers in other countries (U.S. Department of Education, National Center for Education Statistics [U.S. DoE, NCES], 1998). Similarly, many 12th-grade students are doing poorly on the National Assessment of Educational Progress (NAEP). In such tasks and assessments, in almost all cases, U.S. 12th-grade students perform relatively more poorly than 8th-grade students. For example, in TIMSS, 12th-grade

[1] A portion of the study was reported at the Conference on Educational and School Standards, Bad Boll, Germany, 15 December 2003. A revision of the report will be submitted to *Educational Assessment* for publication.

[2] Ann Mastergeorge is now at the Department of Human and Community Development, University of California, Davis.

students are below the international average whereas 8th-grade students are at the international average.

These poor results are usually attributed to cognitive factors such as students' opportunities to learn, teachers' lack of professional preparation, etc. However, a partial explanation of these results may be motivational. Because the low-stakes (for students) tests were administered late in these 12th-graders' final year in high school, the timing of the tests may have negatively affected motivation, and thus performance. This phenomenon has been labeled "senioritis." For the high school senior going into the world of work or on to postsecondary education, tests like TIMSS are clearly low stakes. Thus, one of the major questions about these tests concerns the possible impact of motivational factors on the results. If students are not motivated to perform well on low-stakes tests, then the results may underestimate what students could do if they gave these assessments their best effort.

Rationale

To our knowledge, only two research groups are conducting research using meaningful monetary incentives with released items from NAEP or TIMSS or PISA to test secondary school students: our research group and that headed by Jurgen Baumert at the Max Planck Institute for Human Development in Berlin (see, e.g., Baumert & Demmrich, 2001). In general, a meta-analytic review of research examining the effects of extrinsic rewards on intrinsic motivation (Deci, Koestner, & Ryan, 1999) did not include any studies of this kind. Our basic approach in this research is to provide sufficient monetary incentives to maximize student effort and therefore increase performance. With this approach, we expect that we could stimulate a 0.5 standard deviation increase in performance due to monetary incentives. In our prior study (O'Neil, Sugrue, & Baker, 1996), based on our best-case NAEP data (i.e., $1.00 per test item, on easy items, with 8th graders, for those who remembered their instructions), we found an increase of .41 standard deviation. In that study, we manipulated various incentives (money, task, ego, and standard NAEP instructions) for 8th- and 12th-grade samples of students of different ethnicities (White, Black, Hispanic, and Asian American).

In our current study, because we refined the experimental procedures and offered $10.00 per correct item for 12th graders (or an average of $100 for the testing session), we expected that our monetary incentive would have an effect size as good

as or better than our prior study's effect size of .41. Thus, we predicted (conservatively, we thought) a .5 effect size.

In summary, we had promising results based on our prior NAEP motivation research (O'Neil et al., 1996), and we hypothesized that the new incentive would increase effort, which, along with prior knowledge, would improve performance. The effective incentive in our NAEP study was money. Two issues resulting from that study were controlled for in the current study. First, we expected that the incentive effect might be greater if students believed that they would be rewarded as promised. Some of the participants in the NAEP study were surprised that we actually provided the money. Second, we expected that the incentive effect might be greater when students remembered the treatment group they were in. Our prior study indicated that only approximately two thirds of the students remembered (recognized) what treatment group they were in. We believed that some of the students were not carefully reading the written test instructions. In the current study, we attempted to increase students' beliefs and "remembering" by a combination of (a) a two-item pretest that everyone was expected to answer correctly, followed by immediate payment of $20 cash to the incentive condition participants, and (b) oral delivery of test instructions, using separate rooms for the different treatment conditions, followed by (c) the math literacy assessment.

In general, in our prior study, only the money incentive worked, and only in the 8th grade. The results showed, best case, that the money incentive was effective for a subsample of the 8th-grade students (those who remembered their incentive/treatment group) answering easy and medium difficulty items. With respect to item difficulty results, because the motivational effect occurred at test time, it was not expected that increased effort would improve performance on hard items, because students were not likely to know the content for those items. With respect to remembering their treatment group, presumably if students did not remember the incentive (money), then they would not increase their effort, and thus performance. However, no incentives were effective for 12th-grade students, even those who remembered their treatment group.

We hypothesized that in our prior study, the lack of effect for 12th graders was because (a) the amount of money ($1.00 per item correct) was not large enough to motivate 12th graders, and (b) many 12th graders did not believe they would get the money. Also, we collected the data at the end of the school year (like TIMSS), a time when 12th graders have few reasons to invest effort in low-stakes tests.

Our approach for the current study consisted of manipulating the amount of money per item correct so as to increase the motivational effect and thus increase performance. For our assessment we used the released Third International Mathematics and Science Study (TIMSS) math literacy scale items (TIMSS, 2000). These included both multiple-choice and free-response items. The amount of money given per correct item was either $0 (low-stakes administration, e.g., TIMSS) or $10 per item correct (which we expected to be effective). The incentive group was compared with a group receiving standard, low-stakes TIMSS instructions. Consistent with our prior study, we also collected information on effort, self-efficacy, and worry. The current investigation with 12th graders included a focus group study, a pilot study, a main study, and a supplementary study with Advanced Placement (AP) students in mathematics (the AP study will be reported elsewhere).

Summary of Focus Group Study

The focus group study explored various levels of incentives for 12th graders. The focus group research is documented in detail in Mastergeorge (1999). Parents and students were recruited for participation in the focus groups by teachers at their school sites. A total of eight focus groups (four student groups and four parent groups) were conducted with students and parents from two schools in two different districts in the southern California area. The student groups were composed of sons and daughters of individuals in the parent groups. The AP, high-math-achievement student and parent sample included a total of 12 students and 12 parents in two focus groups of 7 and 5 parents each and two focus groups of 7 and 5 students each; the non-AP, low/medium-math-achievement student and parent sample included a total of 15 students and 15 parents in two focus groups of 8 and 7 parents each and two focus groups of 8 and 7 students each. Although we did not plan originally to include AP students in the pilot or main studies, we included them in the focus group study because we expected these parents and students to be the most knowledgeable about the educational system and the most verbal and vocal regarding possible problems and issues with our study design.

The focus group high schools were chosen for their diverse representation of students across ethnicities, socioeconomic status, and academic performance, and for their participation/nonparticipation in AP courses. Two groups—senior high school students (17- and 18-year-olds) taking AP mathematics and their parents, and

senior high school students (17- and 18-year-olds) taking non-AP courses and their parents—were recruited in order to investigate any similarities or differences among the students and parents regarding their thoughts about incentives on low-stakes tests (Mastergeorge, 1999). The focus groups were conducted to obtain both parent and student perspectives on monetary incentives. A verbal script was read aloud to all participants and included general information about the format, confidentiality issues, and any risk or benefit involved in being a participant in the focus group study (see Mastergeorge, 1999, for the focus group script).

The groups were facilitated by two researchers who engaged participants in discussion in order to ascertain those conditions that might affect students' performance and the amount of money per item that might increase students' motivation to perform on a low-stakes test, and to uncover other variables and parameters that might impact the study (e.g., parental concerns about monetary incentives, security issues, hurt feelings regarding students chosen for incentive versus non-incentive groups, etc.). The following description summarizes the results of the focus group discussions we conducted with parents and students. A more extensive report is provided by Mastergeorge (1999). The text below, from Mastergeorge (1999), gives a flavor of the results. In general, the findings from the focus group study supported our hypotheses (for $10 per item) and allowed us to refine our ideas and procedures.

Parent Focus Groups Discussion: Questions and Answers

1. Suppose your child was given an incentive for getting correct answers on a test. Can you describe/discuss the kinds of "rewards" you would feel comfortable with for correct test items?

   Parents suggested grades, promoting competition between the schools, a year of paid auto insurance (if the student drives), test/class exemptions, extra credit, gift certificates, and tickets to sports games or concerts. Even if parents did not totally agree with the study being done (paying students to perform well), they were comfortable having money as an incentive as long as the students understood that this was a one-time-only study. Many of the parents rewarded their children for good grades by taking them out to dinner, granting them driving privileges, or giving them the chance to make their own decisions, etc., or punished them for not getting good grades by removing driving privileges or "grounding" them. Parents believed the motivation should come from the home, but most seemed to agree that since this was a one-time-only study, they would agree to participate and not be worried about the money and their children's motivation. They felt that cash, checks, and direct deposit would be equally as motivating. Gift certificates would be motivating as well, but most parents thought their

children would prefer money since they would have more choice about what to do with it. They agreed that savings bonds would not be as motivating since the payoff is not as immediate. The parents felt that any amount of money would motivate their teenagers. One group of parents who had children in Advanced Placement classes felt that they would be comfortable with their children receiving $50 at the most, but the rest of the groups felt they would not have a problem with their children receiving as much as $250.

2. Discuss any concerns you might have about your child receiving such an incentive.

   Having students receive cash was not a safety concern (in their schools) especially if they picked it up at the office and the other students did not know how much they got. Of course, it depended on the area in which the study would be done. If the money came in the form of a check, it could be sent to the house.

3. Are there other issues that might affect your child's performance that we should think about related to a "reward?"

   There could be hurt feelings. The students might think it is not fair, or they might feel bad if they do not do well. They should be given a minimum of something for trying—although they should not be told they would be getting it.

Student Focus Groups Discussion: Questions and Answers

1. Suppose you were given an incentive (or reward) for getting correct answers on a test. What kinds of rewards might motivate you to care about getting a correct answer?

   The most popular answer for all of the students was, as expected, money. Many thought other incentives, such as certificates, scholarships, grades, extra credit, class/test exemption, and college recognition, would be motivating as well, but not as motivating as the cash incentive. The problems with gift certificates were that students would need to know before the test where the certificate would be from, and they would have to like the place. The places that were popular would be clothing stores such as Macy's, Old Navy, Footlocker, and the Gap; music stores such as Sam Goody's and Blockbuster; restaurants such as the Olive Garden and TGI Friday's; and movie theatres such as AMC. Many also thought a choice of stores would be a good motivating factor, and the places of choice would depend on the areas that the students lived in. Things like savings bonds would be less motivating because the rewards are not immediate.

   Because many of the students we talked to were planning to go to college next year, money seemed to be the most useful incentive. The college-bound students considered money in the form of direct deposit just as motivating as cash (if they had a bank account), or a check (as long as it was easily

cashed). However, some of the students stated that cashing a check or money order can be a big hassle for them, and often involved a service charge. Amounts as small as $25 ($1 per question correct for a 25-item test) could be motivating; students would try for any money they could get. They felt $5, even $10, per item would be even more motivating, especially if the test was difficult. The value of the amount of money students could get might be influenced by whether they work or not since they would consider how much time they would need to spend on the job in order to get that amount.

2. Do you have any concerns about receiving a reward? Some students taking the test will be in a group without getting a reward, and we want to know if you have concerns about this.

   Most of the students felt that safety was not a concern at their schools. Students often bring money to school and feel safe doing so, because no one really knows how much they have and there is not much of a problem with theft at school. Even the students of lower income backgrounds felt that it would not matter if they received as much as $250 because unless someone knew how much they had, no one would bother them. They could be robbed at any time, whether they had cash, a check, or a money order. Their suggestion was to have the school announce that the participants in the study should go to the principal's office after school and pick up an envelope with the money in it.

3. Are there any other issues or concerns you might have if you were chosen to participate in a test like this?

   There was a concern that some students might feel bad if they tried their best and did not get any right answers. If a student did not get anything from the study, then everyone would know that that student had performed badly. They felt that if a student at least shows up and tries, the student should get something for just participating—even if it is only a small gift certificate. A few felt that the non-incentive group should receive something for participating—of course, they would not be told before they took the test. [We plan to do this.] There was also a concern about unfairness, in that some students may not have been taught the material that is covered on the test. They suggested that the best time of the day to do the study would be in the morning because that is when they will be the most awake, and many students are excused at the end of the day for sports or other extracurricular activities. Most of the students felt their parents would support their participation in the study because they would be getting money that they could use for college. Since the study would be one-time-only, they did not feel that participation in this study would affect their motivation to perform on other tests that have no incentives.

In summary, we were examining the effect of fewer dollars per item correct ($2 to $5 per item correct versus $10 per item correct) in the focus group studies. Thus, we were investigating the magnitude of standard incentives to be used in the main

study. We believed that $10 per test item correct would be appropriate. The $10 figure was also agreed on (instead of $20 per correct test item) at the National Center for Education Statistics (NCES) design review of our study, before we initiated the pilot study. For the focus groups with parents, we were mainly interested in parents' reactions to the incentive idea, security concerns in regard to giving students cash, and whether we should provide payment in the form of checks or certificates. We believed that parental fears were minimal and that because this would be a one-time-only study, there would not be potential opposition. Based on the results of the focus group study, we provided checks as payment.

Summary of the Pilot Study

Although the major purpose of the pilot study was to test the training of assessment administrators and the design of our procedures and forms, we also explored whether or not the treatment (monetary incentive) would increase students' performance in math. It was expected that the mean math score of the students in the incentive group would be higher than the mean score for the control group, and that males would score higher on the math test than females. We consistently find gender effects on math tests with our local urban area samples. We did not expect an interaction between treatment and gender.

Participants

A total of 144 students from five different schools in the southern California area participated in the pilot study. One of the conditions for participating in the pilot study was to be a student in a regular math class. However, 16 students were in AP classes and were dropped from the sample.

Mathematics Test

We used the 20 released math literacy items from the Third International Mathematics and Science Study (TIMSS, 2000). The items ranged in level of difficulty from .26 to .86 (*p* values) based on national norms (Harmon et al., 1997). The items included 12 multiple-choice questions and 8 free-response questions. The multiple-choice items had either four- or five-answer options (see Figure 1 for a multiple-choice item example with correct answer keyed). The free-response items required that the participants show the calculation process, write down an explanation for the response, or draw a graph (see Figure 2 for an item example and Figure 3 for the scoring rubric).

From a batch of 3,000 light bulbs, 100 were selected at random and tested. If 5 of the light bulbs in the sample were found to be defective, how many defective light bulbs would be expected in the entire batch?

A.	*15*
B.	*60*
*C.**	*150*
D.	*300*
E.	*600*

Note. * Correct answer.

Figure 1. Example of multiple-choice item. Source: http://timss.bc.edu/TIMSS1/TIMSSPDF/C_items.pdf

The following two advertisements appeared in a newspaper in a country where the units of currency are zeds.

BUILDING A

Office Space Available

85 - 95 square meters
475 *zeds* per month

100 - 120 square meters
800 *zeds* per month

BUILDING B

Office Space Available

35 - 260 square meters
90 *zeds* per square meter per year

If a company is interested in renting an office of 110 square meters in that country for a year, at which office building, A or B, should they rent the office in order to get the lower price? Show your work.

Figure 2. Example of free-response item. Source: http://timss.bc.edu/TIMSS1/TIMSSPDF/C_items.pdf

Scoring Rubric

Points	Response
Correct response	
2 points	Building A. Correct calculation of rents for both buildings. 9600/800 AND 9900/825, or 825 to compare with 800 given.
2 points	Other correct.
Partial correct	
1 point	Building A. Correct calculation of rent for Building A OR B but not both.
1 point	Building B OR building is not named. Correct calculation of rents for both buildings.
1 point	Building A. Calculations or explanations are incorrect or inadequate.
1 point	Building A. No work shown.
1 point	Building B, OR building is not named. Correct calculation of rent for Building A OR B but not both.
1 point	Building A. Explanation is given only in the form of extracts from the advertisements.
1 point	Other partial.
Incorrect response	
0 points	Building B. Incorrect or inadequate calculations.
0 points	Building B. No work shown.
0 points	Other incorrect.
0 points	Crossed out/erased, illegible, or impossible to interpret.
0 points	BLANK

Figure 3. Scoring rubric for example of free-response item. Source: http://timss.bc.edu/TIMSS1/TIMSSPDF/C_items.pdf

Motivation Questionnaire

In addition to the math test, a state motivation questionnaire was given to participants. This questionnaire (the State Thinking Questionnaire) consisted of three 6-item scales: self-efficacy, worry, and effort. Participants were instructed to indicate how they thought or felt during the math test. The state motivation questionnaire is a modified version of O'Neil, Sugrue, Abedi, Baker, and Golan's (1997) questionnaire, with an added scale for self-efficacy. O'Neil et al. (1997) reported acceptable reliability and validity for these scales.

According to O'Neil and Abedi (1996) and Spielberger (1975), "states" vary in intensity and fluctuate depending on the situation, so the state items used for this study were rated on an intensity dimension with the following responses: *not at all*, *somewhat*, *moderately so,* and *very much so*. These options were scored as 1, 2, 3, and 4 respectively. The directions for completing the questionnaire were as follows:

> A number of statements which people have used to describe themselves are given below. Read each statement and indicate how you thought or felt during the math test. Find the word or phrase that best describes how you thought or felt and circle the number for your answer. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer that seems to describe how you thought or felt during the math test.

An example of a state effort item is "I worked hard on the math test." An example of a state self-efficacy item is "I expected to do very well on the math test." An example of a state worry item is "I was not happy with my performance."

Test Booklets

Two test booklets were created to minimize cheating between students during the test and to reduce the input of item locations on the total test score. The same item set was used with a reversed order of items (e.g., Item 1 in Booklet A was Item 20 in Booklet B, and Item 2 in Booklet A was item 19 in Booklet B, etc.). Booklet A presented a few multiple-choice questions first, followed by a mixture of multiple-choice and free-response questions (this order was the same as the order of the items in the TIMSS released item set). Booklet B presented a few free-response questions first, followed by a mixture of multiple-choice and free-response questions. Equal numbers of Booklets A and B were distributed to students within each classroom.

Procedure

Human Subject Protection Committee approval. Human Subject Protection Committee approval to conduct the investigation for all studies was received from both the University of California, Los Angeles (UCLA), and the University of Southern California (USC). In addition, approval was received from the Committees on Research Studies in the school districts where the studies were conducted.

Test administrators. Test administrators included retired teachers and administrators, and the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) research staff. All test administrators had prior experience with test administration and were further trained for this specific study.

The concept of being paid to do well was explained in the incentive group test administration script, and a reminder phrase was written on the board (i.e., $10/QUESTION). The section of the incentive group test directions that addressed the incentive treatment read as follows:

> Congratulations. This class has been chosen to receive money for each correct answer on this test. We will be giving you each $10 for each correct answer on the math assessment. To show you how it works, we will give you a two-item, very easy test. You will receive $10 in cash for each sample question you get correct today. So, if you get both sample questions correct, you will get $20.
>
> Then we will give you a much harder math test. You will also get $10 per correct item. [WRITE $10/QUESTION on the board.] Since we have to score the math tests, we will get the money to you in 30 days. We will give you the option of receiving a check from UCLA or a post office money order to be sent home once the assessments are corrected.

Test administrators reported anecdotal evidence from the testing sessions that the incentive group participants understood their treatment. This was observed in students' expressions and comments. For example, on hearing about the treatment, many students smiled and some students verbally expressed their excitement. To address the student believability issue, incentive group students received money in class for getting the sample questions correct (see *Random assignment* below). As money was distributed to students for correctly answering the sample questions, many student comments included "Is this for real?" and "You were serious!"

The following directions were read to the control group before taking the test:

> Now turn to the next page titled SECTION 2. Read each question carefully and answer it as well as you can. We will do the two sample questions together in SECTION 2 and you

will complete the other SECTIONS (SECTIONS 3 and 4) on your own. You will be told when to begin each section.

The control group students appeared to accept these directions as familiar, standard directions and made no queries about compensation or payment.

Random assignment. A site coordinator at each school collected signed parental consent forms from the teachers and faxed them to the researchers. We randomly assigned students to either the control group or the incentive group. On the day of testing at each school site, students in participating classes were separated into control, incentive, and nonparticipant groups. We were concerned about the possibility of many seniors deciding not to participate in the study, and thus we publicized the study through the principal, the site coordinator, and classroom teachers. We were successful. Based on a sample consisting of the four initial schools, 80% of the students chose to participate in the study.

Test administration. The test materials were handed out, and the test directions were read aloud to the students. At this point, the incentive group participants were told that they would receive $10 for each correct item on the math test (payment by check, sent through the mail, following the scoring of the tests). To increase believability and thus motivation, the incentive group students were then given two practice questions (see Figure 4) and told that they would receive $10 for each correct answer. The questions were scored by the test administrators and students were paid $10 in cash immediately for each correct answer. All students answered both practice questions correctly. We gave cash on our test days only (maximum of $20 per student). This procedure minimized security concerns. The control group participants were given the same practice questions as in Figure 4, but without a monetary incentive.

Thus the procedures for the control group and the incentive group were the same with the exception that the control group did not receive payment, or the promise of payment, for any of the questions. However, the control group students did receive $20 at the end of the test administration for participating. Students in the control group were unaware during testing that they would receive the $20 upon completion of the test.

Following the practice questions, students in the incentive group and the control group were given 25 minutes to complete the math section and another 8 minutes to complete the motivation questionnaire and the background

SAMPLE QUESTIONS

DIRECTIONS: Read each question carefully and answer it as well as you can.

1. Which of the numbers below is the smallest?

   (A) 3
   (B) * 1
   (C) 4
   (D) 7

2. Which of the numbers below is an even number?

   (A) * 2
   (B) 5
   (C) 3
   (D) 1

Figure 4. Practice questions. Note: Correct answers are starred.

questionnaire. Students were told that they could not go back to any section of the test booklet once the class had moved on to another section. After the test was completed, the test materials were collected for both the incentive and control groups, and students had the opportunity to ask questions regarding the study.

Pilot Study Results

Math Scores

There was no significant difference found between scores of students in the incentive group and students in the control group. Moreover, there was no significant interaction between treatment and gender. However, females (M = 8.50, SD = 3.48, n = 74) had a significantly lower mean math score than males (M = 10.35, SD = .28, n = 54).

Motivation

The motivation questionnaire consisted of three scales of six items each to measure students' levels of effort, self-efficacy, and worry. We first discuss the results for internal consistency of the three scales and then report the relationship between the motivation scales and the math scores.

Internal consistency coefficients for effort, self-efficacy, and worry were .85, .84, and .72 respectively, indicating acceptable reliability for these scales. With respect to effort, there was no impact of treatment or gender. For self-efficacy, the mean for males (M = 16.39, SD = 3.77) was significantly higher than the mean for females (M = 13.71, SD = 3.30). Finally with respect to worry, a significant difference was found only for gender, with males being less worried (M = 11.00, SD = 3.63) than females (M = 12.66, SD = 3.33). Again, there was no treatment effect.

Correlation between math performance and motivation. Correlation coefficients were computed between math score and motivation scale scores in the pilot study to examine the degree of relationship between students' math performance and their effort, self-efficacy, and worry. As expected, there was a significant negative correlation between worry and students' total math score (r = –.43, $p < .01$). There was also a significant correlation between total math score and self-efficacy (r = .42, $p < .01$) and between self-efficacy and effort (r = .35, $p < .01$). Unexpectedly, the correlation between math score and effort was not significant for the total sample.

Summary. The results of the pilot study guided us in the modification of both the instruments and the administration procedures. We made several major modifications to the consent forms and the logistics of test administration for subsequent data collection. Among the most important issues emerging from the pilot testing was the issue that is technically referred to as "diffusion of treatment" (McMillan & Schumacher, 1997). As indicated, there were no significant differences in math performance between the incentive group and the control group. We suspected that in the pilot study, some of the students in the control group may have found out that there was a monetary incentive and thus were motivated to perform better on the math test. The source of this possible contamination was the consent letter and accompanying form that we sent to the parents and the school, as required by the University of California, Los Angeles (UCLA), Human Subject Protection Committee. Parents and students had to sign a consent form in order to participate in the study. The consent form indicated that some students would receive money for each item that they answered correctly. Since students in the control group were tested under the "no money was paid to students" testing condition, learning about the incentive condition may have impacted their performance on the math test. We also made some major modifications to the test instructions and the background questions, beginning with the fourth pilot school site.

In summary, in the pilot study we did not find evidence to support our hypothesis that money would increase students' performance in math. Male students performed better on the math items than female students. There was no significant effect of test booklets. The incentive condition did not increase students' effort or their self-efficacy or worry. However, we felt that with the revised consent forms and procedures, we were ready to test the hypothesis of this investigation in the main study.

Main Study

Hypotheses

We hypothesized that those students receiving $10 per item correct would perform significantly higher in math than those who were not receiving a monetary incentive (the control group). Students receiving an incentive would exhibit higher effort and self-efficacy, but less anxiety, than students in the control group. Our approach consisted of manipulating the testing condition (money or no money per item correct) so as to increase effort and thus increase math performance. In general, we expected overall anxiety levels to be low given the low-stakes nature of the test. Such findings would replicate our prior NAEP findings.

To test the main effects and interaction of treatment and gender, a three-factor completely crossed Analysis of Variance (ANOVA) model was applied to the data. In this model, factor 1 was the treatment effect (incentive versus control), factor 2 was gender (female versus male), and factor 3 was booklet format (A versus B). It was expected that the mean math score of the students in the incentive group would be higher than the mean score of the control group, and that males would perform better on the math test than females. There was no explicit hypothesis for booklet effect, as this variable was used to minimize cheating.

Participants

Four hundred fifteen non-AP students from nine school sites were enrolled in the main study. Students in AP classes were excluded from the main study analyses, first because the admission standards for AP math classes vary dramatically from school to school in the schools' urban location, and second because there are so few AP students in that area. Data were excluded for 22 students enrolled in the main study because those students indicated that they were currently in an AP class or had been enrolled in an AP class (either AP math or AP physics, or both). We did

sample some classes with AP students (N = 21 students) in the main study due to miscommunication with the selected schools. These AP students were also excluded from the main study analyses. However, we conducted additional analyses that included these AP students' data. The results (to be reported elsewhere) suggest that inclusion of these AP students in the analysis does not change the conclusions.

For some of the data analyses (to be reported elsewhere), students were also excluded based on their response to a question asking them to identify which treatment group they were assigned to. The purpose of this question was to identify the issue of treatment as intended versus treatment as remembered. Those participants who could not correctly identify their treatment group were dropped from the analyses. In the main study, 150 students in the incentive group correctly identified that they were to receive money. However, another 5 incentive group students responded that they could not remember. In the control group, 9 students inappropriately responded that they were to receive money; 192 students did not respond to this alternative. However, 35 of these 192 students responded that they could not remember the instructions. An analysis based on data for only those students that remembered their treatment did not substantially change the results.

Three students received booklets that did not contain all math test questions; data for those students were excluded from the analyses. One student marked "1" for all questions on one section of the motivation questionnaire, and that student's responses were excluded from the motivation part of the analyses. In the main study, due to an unanticipated increase in the number of eligible participants in four classrooms, calculators were not available for 41 students. Data for these students were therefore excluded in part of the analyses.

Materials

We used the same math test, motivation questionnaire, and modified background questionnaire as in the pilot study. We also collected additional information on students' math achievement level, language background, opportunity to learn, and scores on the Stanford Achievement Test, 9th edition (SAT-9). These data are reported in O'Neil, Abedi, Lee, Miyoshi, and Mastergeorge (2000).

The SAT-9 is taken by all K-11 students in the districts we sampled. Thus, we considered a school site's SAT-9 national percentile rank in math to be a better indicator of overall school performance in math. Therefore, in the main study, SAT-9

scores were used in the site selection process. These math percentile scores ranged from 28 to 50 with a median percentile math National Percentile Rank of 36. Of the schools that agreed to participate in our study, the best performing school had a National Percentile Rank of 50 on the 1998 math SAT-9. Most of the school sites selected for the main study were in the medium to low range for performance level on the SAT-9. An effort was made to recruit school sites that were in the high range on the SAT-9; those sites that were contacted declined to participate.

Analyses conducted during the main study revealed that the control group and the incentive group did not differ significantly in math performance. As with the consent form in the pilot study, a review of the letter introducing the study to the principal revealed that a monetary incentive was discussed. We suspected that the principal and teachers at some school sites may have revealed information about the incentive money to their students. We therefore revised the principal's letter and recruited a new group of school sites in a single district for participation in the study. These sites received the revised principal's letter and the revised consent form.

Thus, the main study sites comprised five schools that received the original principal's letter and revised consent form and four schools that received the revised principal's letter and revised consent form. Analyses were performed to determine whether differences existed within these groups. We categorized schools into three groups: Group 1, schools receiving the original principal's letter and original consent form; Group 2, schools receiving the original principal's letter and revised consent form; and Group 3, schools receiving the revised principal's letter and revised consent form. The mean math score for Group 1 (original letter and original consent) was 10.31 (SD = 4.81, n = 144); for Group 2, the mean was 7.84 (SD = 4.00, n = 238); and for Group 3, the mean was 8.53 (SD = 4.00, n = 177).

To test the performance of the three groups of schools across the categories of treatment (treatment, control), a two-factor ANOVA model was used. Mean differences between the incentive and control groups for all schools were not significant. Mean differences between the three groups of schools were significant, $F(2, 553) = 15.69$, $p < .001$. To compare the means of the three groups, we used the Tukey Honest Significant Difference (HSD) multiple comparison approach. The results of analyses indicated that the means for Group 1 schools were significantly different from the means for Group 2 schools (HSD = 2.48, $p < .01$) and Group 3 schools (HSD = 1.74, $p <. 01$), and that the Group 2 mean was not significantly

different from the mean for Group 3 (HSD = -.69, p = .228). The interaction between the school groups and treatment was not significant. Thus, the school factor (the three groups) had no statistical effect as a function of treatment.

Main Study Analyses

For power analyses and computation of sample size, we used data from our current pilot study and from the earlier CRESST motivation studies. We estimated the number of participants needed to detect our hypothesized difference of a .5 standard deviation difference on the math test.

Background questionnaire. Based on the main study participants' self-reported background information in regard to the types of math and physics classes taken, 13 students in the incentive group and 8 students in the control group were enrolled in AP math classes, and 1 student was enrolled in an AP physics class. These 22 students were excluded from all of the analyses in the main study.

In the background questionnaire, we asked students whether they spoke a language other than English at home, and if they did, how often they used that language. The question had three responses: *always, sometimes*, and *rarely*. Twenty-three students from the incentive group and 48 students from the control group reported that they never used a language other than English at home. Fifty-eight students from the incentive group and 121 students from the control group reported that they sometimes used a language other than English at home. One hundred ten students from the incentive group and 212 students from the control group reported that they always used a language other than English at home.

We used a one-factor ANOVA model to compare students' math performance across the categories of this variable. The mean score for the group of students responding *always* was 8.53 (SD = 4.66, n = 48); for the group responding *sometimes*, the mean was 7.29 (SD = 3.59, n = 121); and for the group responding *rarely*, the mean was 8.31 (SD = 3.68, n = 212). The results of analyses of variance showed the difference between the three groups to be significant. The significant difference is mainly due to the difference between the group of students who responded *sometimes* and the group who responded *rarely*. Tukey HDS tests showed only one significant difference, between the means of the groups responding *sometimes* and *rarely*.

The main research hypotheses focus on the differences between the math performance of the incentive and control groups. We included gender and booklet

as two additional independent variables in this study. Thus, a three-factor completely crossed ANOVA model was applied to the data. Table 1 shows the means and standard deviations for students in the incentive and control groups by gender and booklet. The range of possible scores was 0-24 (there were 20 items but some items were scored 0-2 points).

As the data in Table 1 show, the students in the incentive group (M = 7.97, SD = 3.73) performed no better than the students in the control group (M = 7.94, SD = 3.86). However, males had significantly higher mean math scores (M = 8.81, SD = 4.06) than females (M = 7.11, SD = 3.31). In addition, booklet format appeared to make a difference. Students who received Booklet B (M = 8.39, SD = 3.73) had higher scores than students who received Booklet A (M = 7.47, SD = 3.81), and this difference was significant. The treatment by gender by booklet interaction was also statistically significant. For Booklet A, males in the incentive group outperformed males in the control group (M = 9.13 vs. M = 7.91), whereas mean scores for females

Table 1

Descriptive Statistics for Math Test Score by Treatment, Gender, and Booklet for the Main Study Sample

	Treatment								
	Incentive			Control			Total		
	Male	Female	Total	Male	Female	Total	Male	Female	Total
Booklet A									
M	9.13	6.33	7.62	7.91	6.84	7.32	8.51	6.59	7.47
SD	4.34	3.55	4.15	3.81	3.10	3.46	4.10	3.32	3.81
n	42	49	91	43	52	95	85	101	186
Booklet B									
M	8.46	8.07	8.28	9.58	7.19	8.51	9.04	7.65	8.39
SD	3.53	3.08	3.31	4.42	3.33	4.13	4.03	3.22	3.73
n	54	50	104	57	46	103	111	96	207
Total									
M	8.75	7.21	7.97	8.86	7.00	7.94	8.81	7.11	7.96
SD	3.90	3.42	3.73	4.23	3.20	3.86	4.06	3.31	3.79
n	96	99	195	100	98	198	196	197	393

Note. For questions with mutually exclusive categories, totals may not add up due to missing data. For those that are not mutually exclusive, the total number of responses may be larger than the total number of participants due to the possibility of multiple selective categories. There were 20 questions with a possible score range of 0-24 points; some items were scored 0-2 points.

in the incentive and control groups were similar (M = 6.33 vs. M = 6.84 respectively). However, for Booklet B, males in the control group outperformed males in the incentive group (M = 9.58 vs. M = 8.46), whereas females in the incentive group outperformed females in the control group (M = 8.07 vs. M = 7.19). Yet, post hoc analyses of these mean differences indicated that they were not significantly different. Although the overall interaction was significant, these means were not significantly different from each other. The results of the Tukey HSD test showed that for Booklet A, males in the incentive group scored higher than females in the incentive group, whereas for Booklet B, there was no significant difference in mean scores of incentive group males and incentive group females. Also, for Booklet A, males in the control group did not score differently from females in the control group, but for Booklet B, control group males outperformed control group females.

A two-factor analysis of covariance (ANCOVA) design was used to test the main and interaction effects of treatment and gender on math when students' reading performance was controlled for. Thus, SAT-9 reading score was used as a covariate. For the main study, the mean math score for the incentive group was 7.72 (SD = 3.73, n = 62) and for the control group, the mean was 7.62 (SD = 4.02, n = 61). The means for the two groups were very similar and not significantly different. For males, the mean was 8.42 (SD = 3.84, n = 68), and for females, the mean was 6.75 (SD = 3.72, n = 55). The difference between the performance of males and females was significant. The interaction between treatment and gender was not significant. The smaller number of participants in this design was caused by missing data on the SAT-9 reading scores. The adjusted means are: for males, M = 8.42 (SE = .45); for females, M = 6.76 (SE = .48); for the control group, M = 7.53 (SE = .48); and for the incentive group, M = 7.65 (SE = .48).

Item difficulty. The motivational effect was investigated at test time, so it was not expected that increased effort would improve performance on difficult items, because students were unlikely to be familiar with the content of those items. We expected (as in our prior study, O'Neil et al., 1996) that the motivation effect at test time would be most salient on easy items, as "easy" would indicate prior knowledge, and thus incentives could lead to more effort, and more effort with prior knowledge could lead to higher math performance. Based on the TIMSS item p values (proportion of item correct response), which we obtained from the TIMSS assessment (Harmon et al., 1997), subsets of TIMSS test items were used to create two test scores: (a) for easy items, and (b) for difficult items. Five items (Questions 2,

8, 10, 16, and 17, percent correct > .64) were considered easy items (see O'Neil et al., 2000, Appendix F, Booklet A only). Five items (Questions 4, 5, 13, 19, and 20, percent correct < .28) were considered difficult items. The mean for the five easy items was 3.54; for the five difficult items, the mean was 0.9., a substantial difference. The maximum possible score for the five easy items was 5 points. The maximum possible score for the five difficult items was 8 points. The easy and difficult test scores were used successively in a $2 \times 2 \times 2$ completely crossed ANOVA model, which we applied to the total math scores. Item type (easy and difficult composite test scores) was used as the within-subject factor, and treatment, gender, and booklet were used as the between-subject factors.

Easy items. The overall mean score for easy items in the main study was 2.95 (*SD* = 1.26). Thus, for the easy items (based on national norms), the percent correct for the main study sample was 59%, which in our sample would not signify "easy." The mean score for the five easy items for the incentive group was 2.97 (*SD* = 1.25), and for the control group it was 2.93 (*SD* = 1.27). There was no main effect of treatment. There was, however, a significant gender difference on the easy items: The mean score for males was 3.27 (*SD* = 1.16), whereas for females, the mean score was 2.63 (*SD* = 1.28). Students who used Booklet B (*M* = 3.13, *SD* = 1.11) performed significantly better than those who used Booklet A (*M* = 2.74, *SD* = 1.38). The interactions among treatment, gender, and booklet were not statistically significant.

Difficult items. The mean score for difficult items was 1.42 (*SD* = 1.38). The maximum possible score for the five difficult items was 8 points. Thus, for the difficult items, the percent correct for the main study sample was 18%, indicating a very difficult set of items. There was no treatment main effect for the incentive group (*M* = 1.45, *SD* = 1.36) compared with the control group (*M* = 1.39, *SD* = 1.39). There was a significant difference between scores for males (*M* = 1.61, *SD* = 1.50) and females (*M* = 1.23, *SD* = 1.21) on the difficult items. There was no booklet effect. Finally, none of the interactions was significant.

Motivation

Internal consistency coefficients were computed for the three motivation scales, which showed a high level of internal consistency with alpha coefficients of .85 for effort, .84 for self-efficacy, and .72 for worry.

To compare students' responses across categories of treatment (incentive / control), gender (male / female), and booklet (A / B), a $2 \times 2 \times 2$ ANOVA

model was used. Scores for the three motivation scales (effort, self-efficacy, and worry) were used as the dependent variables in separate ANOVAs.

Effort. The overall mean score for effort for the main study was 18.09 (SD = 4.15) out of a possible 24 points, indicating that the students in the main study exhibited moderate effort. The mean effort scores for females (M = 18.25, SD = 3.30) and for males (M = 17.92, SD = 4.48) were almost identical, and the difference was not significant. There was a significant difference between the levels of effort across the treatment groups. The mean effort score for the incentive group was 19.17 (SD = 3.70), and for the control group, the mean was 17.00 (SD = 4.30). This difference of about 2 score points is significant, which indicates that the incentive group put more effort into this test. Booklet form did not have a significant effect on effort. The interactions were not significant.

Self-efficacy. The overall mean score for self-efficacy was 14.66 (SD = 3.90) from a maximum of 24 possible points, indicating low self-efficacy. The mean score for the incentive group was 15.23 (SD = 3.87), which is significantly higher than the mean score of 14.09 (SD = 3.86) for the control group. The results also showed a significant gender difference. The mean self-efficacy score for males (M = 15.72, SD = 3.82) was significantly higher than the mean score for the females (M = 13.66, SD = 3.72). Booklet form had no significant impact on self-efficacy; for students who used Booklet A, the mean was 14.29 (SD = 4.10), and for students who used Booklet B, the mean was 14.99 (SD = 3.69). None of the interactions was significant.

Worry. The overall mean worry score for the main study sample was 12.21 (SD = 3.96) from a maximum of 24 points, indicating very low worry. The mean worry scores for the incentive group (M = 12.47, SD = 4.16) and for the control group (M = 11.94, SD = 3.74) were approximately equal. The mean worry score for females was 12.65 (SD = 3.86), and for males, the mean score was 11.73 (SD = 4.02). This difference was significant. Booklet form also had a significant impact on the worry level. The mean worry score for students who used Booklet A was 12.75 (SD = 3.78), and for students who used Booklet B, it was 11.22 (SD = 3.58). Given that Booklet B was easier, these results are consistent as worry tracks task difficulty.

Relationship between math performance and motivation. Table 2 presents the set of correlations between math performance and motivation. There was no significant relationship between level of effort and math performance, but the other expected relationships were significant (e.g., more worry / poorer performance). A comparison of the treatment and control groups separately can be found in Table 3.

Table 2

Correlation Coefficients Between Math Test Scores and Motivation Scale Scores for the Total Main Study Sample

		Total math	Effort	Self-efficacy	Worry
Total math	*r*	1.00			
	n	393			
Effort	*r*	.10	1.00		
	n	382	382		
Self-efficacy	*r*	.40**	.44**	1.00	
	n	376	395	396	
Worry	*r*	–.34**	.14*	–.26**	1.00
	n	378	376	370	378

*$p < .05$, two-tailed. **$p < .01$, two-tailed.

Table 3

Comparison of Correlation Relationships Between Motivation and Math Performance for the Main Study Sample by Treatment

		Incentive	Control
Effort / Math performance		.02	.12
	n	205	199
Self-efficacy / Math performance		.38**	.45**
	n	201	196
Worry / Math performance		–.38**	–.38**
	n	203	196

**$p < .01$, two-tailed.

Analysis of the omitted/not-reached items. Another measure of motivation is the number of omitted and not-reached items (see Table 4). The mean number of math items that were omitted and the mean number of items that were not reached were obtained. Omitted items are defined as those items that are left blank and are followed by some attempted items. Not-reached items are those that are left blank and are followed by no attempted items. We hypothesized that the incentive condition would increase effort and that such higher effort would result in fewer omitted and not-reached items.

Table 4

Frequency Distribution of the Omitted and Not-Reached Items by Treatment

Group		Incentive	Control
Incentive		1.45	2.69
	n	195	195
Control		1.11	2.42
	n	198	198

The mean number of not-reached items was used as the dependent variable in a two-factor ANOVA in which gender and treatment were the two independent variables. Because of the small *n*s in some cells, a 2 x 2 x 2 design was not used. The results of analyses on not-reached items showed no significant main effects or interactions. In analysis on the omitted items, using the same design, however, treatment effect was significant, $F = 8.23$, $df = 1, 555$, $p = < .001$. The incentive group omitted a larger number of items than the control group. The mean number of items omitted for the incentive group was 1.38 as compared with a mean of 1.01 for the control group.

Discussion of the Main Study

Recent information in the 1990s on international assessments (e.g., the Third International Mathematics and Science Study [TIMSS]) indicates that 12th-grade students in the United States are doing extremely poorly on such assessments compared with their peers in other countries (U.S. DoE, NCES, 1998). Similarly, many 12th-grade students are doing poorly on the National Assessment of Educational Progress (NAEP). In such tasks and assessments, in almost all cases, U.S. 12th-grade students performed relatively worse than 8th-grade students. For example, in TIMSS, 12th-grade students were below the international average whereas 8th-grade students were at the international average. Similar results are reported on TIMSS-R (Martin, Gregory, & Stemler, 2000; Martin, Mullis, et al., 2000; Mullis et al., 2000).

These poor results are usually attributed to cognitive factors such as students' opportunity to learn, teachers' lack of professional preparation, etc. However, a partial explanation of these results may be motivational. Because the low-stakes (for students) tests were administered late in these 12th-graders' final year in high

school, the timing may have negatively affected motivation, and thus performance. This phenomenon has been labeled "senioritis." For the high school senior going into the world of work or on to postsecondary education, tests like TIMSS are clearly low stakes. Thus, one of the major questions about these tests concerns the possible impact of motivational factors on the results. If students are not motivated to perform well on low-stakes tests, then the results may underestimate what students could do if they gave these assessments their best effort.

Our basic approach was to provide a sufficient monetary incentive to maximize student effort and therefore increase performance. We expected that we could stimulate a 0.5 standard deviation increase in performance due to such incentives. Our results will not generalize, without additional research, to either TIMSS or NAEP. Further, our results will not generalize to the impact of motivation variables (e.g., effort, self-efficacy) on the teaching and learning of math. However, we expected our results to constitute a proof of concept of the importance of manipulating motivation in low-stakes assessments for 12th graders.

We had promising results based on our prior NAEP motivation research sponsored by the National Assessment Governing Board (NAGB), Office of Educational Research and Improvement (OERI)/National Center for Education Statistics (NCES). We hypothesized that the incentives would increase effort, which along with prior knowledge would improve performance. The effective incentive in this earlier study (O'Neil, Sugrue, Abedi, Baker, & Golan, 1992) was money. In the study, we manipulated various incentives (money, task, ego, standard NAEP instructions) for 8th- and 12th-grade samples of students of different ethnicities (White, Black, Hispanic, and Asian American).

In general, only the money incentive worked and only in the 8th grade. The results showed, in the best case, that the money incentive was effective for a subsample of the 8th-grade students (those who remembered their incentive/treatment group) tested on easy and medium difficulty items. With respect to item difficulty results, because the incentive, and therefore the motivational effect, was available only at test time, it was not expected that increased effort would improve performance on hard items, because students were unlikely to know the content. With respect to remembering their treatment group, presumably if students did not remember the incentive (money), then they would not increase their effort, and thus performance. However, no incentives were

effective for 12th-grade students, even for those who remembered their treatment group.

We hypothesized that in our prior study, the lack of effect for 12th graders was because (a) the amount of money ($1.00 per item correct) was not large enough to motivate 12th graders, and (b) many 12th graders did not believe they would get the money.

Our approach in the current study consisted of manipulating the amount of money per item correct so as to increase the motivational effect and thus increase performance. The amount of money given per correct item was either $0 (low-stakes administration, e.g., TIMSS) or $10 per item correct (which we expected to be effective). The incentive group was compared with a group receiving standard low-stakes TIMSS instructions. Consistent with our prior study, we also collected information on effort, self-efficacy, and worry. For our assessment we used the released TIMSS math literacy scale items, which included both multiple-choice and free-response items.

We hypothesized that students receiving $10 per item correct (incentive group) would perform significantly higher on the math assessment than those who were not receiving any monetary incentive (control group). Students in the incentive group were also expected to exhibit more effort and self-efficacy and less worry than control group students.

This investigation with 12th graders included a focus group study, a pilot study, a main study, and a supplementary study (reported elsewhere) with AP students in mathematics. In the focus group study (documented in Mastergeorge, 1999) we explored various levels of incentives. Parents and students who participated in the focus groups suggested that $5 to $10 per item correct would provide enough motivation for students in Grade 12 to work harder on math test items. Based on these findings, we offered students $10 per item correct in the present investigation to find out whether their performance on the selected math items could be increased under such a high-stakes testing condition. We then compared the performance of students receiving $10 per item correct with the performance of students who responded to the same set of items with no monetary incentive.

A total of 559 students participated in the pilot and main studies (144 students in the pilot study, and 415 students in the main study). For the pilot and main

studies, students were selected from 14 different schools (5 schools in the pilot study, 9 schools in the main study) in southern California school districts. These schools had different demographic profiles and different levels of overall student performance.

Following the focus group study, we conducted a pilot study. The purpose of the pilot study was to test design issues, examine the accuracy and language of the instruments, and resolve logistical problems. The results of the pilot study helped us to refine the instruments and to modify the design. We then conducted the main study.

For an approximately 1-hour testing session, the average student in the incentive condition in the main study received $100 ($80 for an average of 7.96 items correctly answered and $20 for the two "easy" practice test items). Such incentives were assumed to be motivational for the 12th graders in our samples. However, the results of the main study showed no significant difference between the performance of students in the incentive and control groups. Statistically, there was no main effect of the incentive treatment. In the main study there was a complex interaction between treatment, gender, and booklet. However, post hoc comparisons indicated that although the overall interaction was significant, none of the comparisons of appropriate means were statistically significant. Thus, we chose to be conservative and not to interpret this interaction as supporting our major hypothesis. The total number of students in the main study was 393 after excluding students with incomplete data, and when participants were divided into subgroups by independent variables such as gender, test form, and treatment, the number of students in each group was smaller yet. Thus, due to the small numbers of students, for some of the analyses there was not enough power to detect a significant difference, even when the difference was relatively large. However there were a sufficient number of students in the main study sample to detect a reasonable main effect for the incentive treatment.

There was a great deal of consistency in the data in the main study. For example, males performed significantly better than females. These results were expected as the task was mathematics, and although in the national sample (U.S. DoE, NCES, 1998) there were no significant effects of gender for mathematics, we consistently find gender effects on math tests with our local southern California samples. In the main study, students reported significantly more effort in the incentive condition than in the control condition. Finally, in the main study, self-

efficacy and effort were positively related. These latter results make theoretical sense, as Bandura (1986, 1993, 1997) predicted that higher levels of self-efficacy should lead to higher levels of effort.

We also predicted, based on our prior NAEP research, that the incentive condition should result in higher effort. In the main study we found that students in the incentive group reported significantly higher effort than students did in the control group. In turn, this increased effort should have resulted in better math performance. So why did we not find a significant main effect of treatment on math performance, given that there was a main effect of treatment on effort? The obvious explanations (e.g., poor reliability of the measures) are not true. The alpha reliability of the effort scale was .85 in the main study. Further, the correlations between effort and self-efficacy and worry were significant in the predicted directions for the main study, indicating that other validity predictions involving effort were consistent with our prior research and the literature.

The major reason, we felt, was the lack of correlation between self-reported effort and math achievement. Unexpectedly, for the main study, self-reported effort was not significantly related to math performance. With respect to effort, the research literature and our own prior research using the same measures indicated that the relationship would be positive (i.e., higher effort leads to better performance). Not surprisingly, we are puzzled by such findings. There was no issue of whether enough time was provided to complete the math test, given the number of not-reached items was very low, which indicates that students had sufficient time to complete almost all items on the test. Further, there were few items omitted. Information on the not-reached and omitted items indicates that students had sufficient time to complete the test. Thus our set of items clearly constituted a power test, not a speed test. Further, for the total math items correct, there was no ceiling. In the main study, the mean was 7.96 (SD = 3.79) out of a possible 24 points (20 items, with a few extended response items getting 2 possible maximum points).

We also had several other behavioral indicators that the students put effort into other aspects of task performance, for example, an indicator based on the number of checks cashed by incentive participants. Because we needed time to score the performance items, and to minimize security concerns (cash in the hands of 12th graders), we asked the incentive group students, before the math test, to complete a form indicating where we should send the money they would receive for performing successfully on the math test. For the pilot study and the main study

combined, 279 participants requested a check (one student requested a money order), and 272 students' checks (or 98%) cleared the bank. (We do not know why 6 students did not cash their checks.) Thus, it appears that students in the incentive condition were motivated to expend effort to correctly fill out the form in the student test booklet to obtain the money.

Other behavioral information from the main study seems to indicate that our oral and written instructions resulted in students paying attention to the instructions for the math test in general. For example, the two "easy items" were completed without error by all incentive and control group participants, indicating that the experimental controls (e.g., for believability) were effective. An interesting finding was that the incentive group had significantly fewer omitted items; thus, their strategy seems to have been to attempt fewer items (not more, as we expected) and take time to make sure that they would answer those items correctly. Given that there was no significant difference in total math performance between the two groups, the incentive group strategy resulted in fewer items attempted but greater success on those items.

There is an additional issue that speaks to our location in southern California. One way in which the sample of students in the current study is not representative of all students in the United States is that so many students in this study are from families that do not speak English at home. Well over half of the students in the study indicated that they never speak English at home, whereas less than 15% indicated speaking only English at home. We analyzed math performance as a function of this variable and found no relationship with treatment.

The findings are the same for the analyses using subsets of the items (i.e., the "easy" and the "difficult" items). Moreover, we ran additional analyses with AP students included in the main sample study. The results were exactly the same for math as when the AP students were excluded. Thus, we feel that, although troubling, the finding that 30% of the students could not remember that they were to receive money does not affect our conclusions.

In our prior study (O'Neil et al., 1992, 1997), in which a money incentive was not motivating for 12th graders, we hypothesized that the lack of effect for 12th graders was because (a) the amount of money ($1.00 per item correct) was not large enough for 12th graders, and (b) many 12th graders did not believe they would get the money. By comparison, in the present study, we felt that both conditions (i.e.,

amount of money and believability) were satisfactory and should have been motivating, but they were not. As in our prior study, our chain of logic was that money would be an incentive to increase effort and therefore improve math performance. We succeeded in increasing effort (measured by self-report and the number of checks that cleared), but the incentive condition did not improve math performance. The mechanism of high effort leading to better math performance for those with prior knowledge at test time was based on our prior research, the literature, and common sense.

Presumably, since the incentive condition increased effort in the main study, if self-reported effort were related to performance, then the incentive condition would have increased math performance. One might argue that there was a suggestion of an incentive effect in the significant triple interaction between incentive, gender, and test booklet in the main study. However, this effect was relatively weak, as post hoc comparisons indicated no significant difference for the mean comparisons. Thus, as mentioned earlier, we discounted this interaction.

In summary, effort was not related to performance, and the conclusion for this set of studies is that a strong monetary incentive did not increase math performance on a set of TIMSS released math items with a local sample including a large proportion of English language learners from samples of convenience. Similar findings for a German sample were reported by Baumert and Demmrich (2001). Further, the inability to find motivational effects, despite a strong incentive, random assignment (with equivalence on background characteristics), tests of high- and low-performing students, and elimination of non-accurate recall cases, is quite compelling. It raises some fundamental questions about previous assumptions made about the motivation effect on test performance and, we think, allows large-scale, low-stakes assessments to move forward with more confidence about the integrity of their results. We believe that there is a senioritis effect, but understanding its specific motivational effect on test performance and its amelioration await future research. The obvious next step would be to conduct a series of focus groups and to design cognitive laboratory approaches to better understand these issues.

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28,* 117-148.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

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

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

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., et al. (1997, September). *Performance assessment in IEA's Third International Mathematics And Science Study (TIMSS).* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 technical report.* Chestnut Hill, MA: Boston College, Lynch School of Education, International Study Center.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., et al. (2000). *TIMSS 1999 international science report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill, MA: Boston College, Lynch School of Education, International Study Center.

Mastergeorge, A. (1999). *Focus groups on motivational incentives for low-stakes tests with senior high school students and their parents* (Report to AIR/ESSI). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

McMillan, J. H., & Schumacher, S. (1997). *Research in education* (4th ed.). New York: Longman.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., et al. (2000). *TIMSS 1999 international mathematics report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth grade.* Chestnut Hill, MA: Boston College, Lynch School of Education, International Study Center.

O'Neil, H. F., Jr., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *Journal of Educational Research, 89,* 234-245.

O'Neil, H. F., Jr., Abedi, J., Lee, C.-L., Miyoshi, J., & Mastergeorge, A. (2000). *Motivational incentives for low-stakes tests* (Draft final report to NCES). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). *Final report of experimental studies on motivation and NAEP test performance* (Report to NCES). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). *Final report of experimental studies on motivation and NAEP test performance* (CSE Tech. Rep. No. 427). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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

Spielberger, C. D. (1975). Anxiety: State-trait process. In C. D. Spielberger & I. G. Sarason (Eds.), *Stress and anxiety* (Vol. 1, pp. 115-143). Washington, DC: Hemisphere.

Third International Mathematics and Science Study. (2000). *TIMSS released item set for the final year of secondary school.* Available 11 July 2000, http://timss.bc.edu/TIMSS1/TIMSSPDF/C_items.pdf

U.S. Department of Education, National Center for Education Statistics. (1998). *Pursuing excellence: A study of U.S. twelfth-grade mathematics and science achievement in international context* (Rep. No. NCES 98-049). Washington, DC: Author.