659: DEGREE PROJECT IN ECONOMICS
BACHELOR'S THESIS
DEPARTMENT OF ECONOMICS
STOCKHOLM SCHOOL OF ECONOMICS
SPRING 2013

INCENTIVES AFFECTING TEST-TAKING BEHAVIOR: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN SWEDISH ELEMENTARY SCHOOLS

Nina Jalava
22334@student.hhs.se

Elin Pellas
22299@student.hhs.se

Abstract

This study examines the effect of non-financial incentives on elementary school student performance. An experiment on sixth graders in Swedish elementary schools was conducted to measure the effect of different incentives applied in the test situation. Our results indicate that extrinsic incentives play an important role in motivating students to exert more effort. We observe significant differences in the means of test scores between the control group and three out of four treatment groups. The treatments evaluated are: (i) students receiving grades A-F, (ii) students receiving grade A if they are among the top 3 performing students, (iii) students receiving a diploma if they obtain a pre-determined score or above, and (iv) students receiving a prize if they are among the top 3 performing students. Treatments (ii)-(iv) represent alternative ways of motivating students. We find that the only treatment with insignificant difference in means is (i) students receiving grades A-F. This assessment method is the one traditionally being applied. Furthermore, we find that motivational strengths of evaluated incentives differ with respect to gender. Our results call into question the current structure of the educational system in motivating students in test situations.

JEL: I20, I21, C90, D03
Keywords: Test-taking, Examinee Effort, Performance Incentives, Extrinsic Motivation, Randomized Experiments

DATE: May 29, 2013
PLACE: Stockholm School of Economics, Sveavägen 65
EXAMINER: Erik Meyersson
DISCUSSANTS: My Hedlin and Max Reppen
SUPERVISOR: Juanna Schrøter Joensen

Acknowledgments

We would like to express our gratitude to the teachers and students who made this study possible. In no particular order, we wish to thank the following teachers and classes for their time and collaboration: Maud Larsson-Tegelgård and class 6A and 6B at Hägerstensåsens skola; Anna Larsson, Lars Edlund and class 6A and 6B at Gubbängsskolan; Åsa Eriksson, Fredrik Vestin and class 6A and 6E at Mälarhöjdens skola; Hans Flygård and class 6:1, 6:2 and 6:3 at Bagarmossens skola; Rima Mikayel, Heba Khalaf, Cattis and class 6A, 6B and 6C at Hässelbygårdsskolan; Håkan Karlsson and class 6E and 6F at Adolf Fredriks Musikklasser; Krister Werner, Roxana Talaremi and class 6A, 6B and 6C at Katarina Norra skola; Clas Gordon and class 6A and 6B at Katarina Södra skola; Catarina Schulz, Jennie Thulin and class 6 at Fredrikshovs Slotts skola; Åsa Rikardsson-Fundin, Johan and class 6A and 6B at Kungsholmens grundskola. Visiting you has been a very inspiring and personally rewarding experience.

We also wish to express our appreciation for all the support we received from our supervisor Assistant Professor Juanna Schrøter Joensen at the Department of Economics at the Stockholm School of Economics. Thank you for your curiosity and helpful feedback throughout the process. It enabled us to proceed in the right direction.

Finally, we wish to acknowledge the help we received from our friends and family. Special thanks to José Araújo, Niklas Jalava, Carola Pellas, Olav Pellas and Sakiko Reuterskiöld. Your comments and suggestions were greatly appreciated.

Contents

List of Figures

List of Tables

1 Introduction

Student performance goes hand in hand with student motivation and effort. Improving student performance is a key issue in most educational discussions and much time and resources are devoted to this aim. Most often, student quality is determined by the performance on various tests, yet little research has been conducted examining student motivation in test situations. Studying motivation in an educational environment to better understand how students respond to different test setups and to identify how students react to the application of different incentives can benefit the educational system. In this paper, we conduct an experiment to study the effect of non-financial methods of incentivizing elementary school students.

Grading students is used as a screening device in school admission procedures and is also thought to ensure higher student performance. Continuous grading means that a student's results are effectively communicated between schools and families and that those students who require additional assistance are identified and can receive necessary support. But what are the short-term consequences of grading on student motivation and effort? How do students respond to being told that they will be graded, particularly on a test that is low-stake for the students but can carry high stakes for the teachers and schools administrating the test? And what happens when we introduce other ways of incentivizing students?

In this study, we analyze the effects of receiving extrinsic incentives on student effort. An experiment on 437 elementary school students is conducted to evaluate how short term effort can be affected by students receiving different information concerning the assessment of the test they are taking. We are interested in evaluating non-financial means of incentivizing students since these are the most cost-effective for schools to use and implement. Additionally, previous research has shown that elementary school students respond well to non-financial incentives (Levitt et al., 2012). The specific incentives we

are analyzing are (i) students receiving grades A-F, (ii) students receiving grade A if they are among the top 3 performing students, (iii) students receiving a diploma if they exceed a pre-determined score, and (iv) students receiving a prize if they are among the top 3 performing students.

Our experiment is conducted on Swedish elementary school students, and was chosen to include sixth graders because Sweden has recently implemented grading for students in the sixth grade. Previously, students in Sweden received grades only once they reached eighth grade of middle school. Swedish students have also been shown to consistently underperform compared to neighboring countries in the Programme for International Student Assessment (PISA) study conducted by the Organisation for Economic Co-operation and Development (OECD) (Margaret and Ray, 2003). In 2009, Sweden placed 28th overall of a total of 65 participating countries, with an average total score of 496 points. This can be compared to the OECD average score of 497, and neighboring countries' results; Finland obtained an average score of 543, Norway 500, and Denmark 499. Sweden's results have also worsened since the previous PISA study in 2006, in which Sweden obtained results above the OECD average (Ekonomifakta, 2009). Consequently, ways to achieve higher student performance in Swedish schools has received considerable attention in educational debate.

By approaching the issue of student motivation and performance from a behavioral economics perspective, we may be able to help achieve a better comprehension of student motivation in test situations and better understand how students react to different incentives.

2 Previous research and theory

2.1 Intrinsic and extrinsic motivation

The distinction between intrinsic and extrinsic motivation is fundamental in an educational setting. Intrinsic motivation refers to motivation coming from within the subjects themselves and is driven by an interest in, or enjoyment of, the task itself. Extrinsic motivation relies on external factors as a driving force for motivating the subjects. Many studies on motivating students extrinsically have used financial means as a method of motivating; paying students has been shown to result in better performance.(Eisenkopf, 2011; Fryer, 2011; Bettinger and Slonim, 2007). There is a worry that the introduction of extrinsic incentives can have a detrimental effect on students' future performance, as extrinsic motivation may crowd out intrinsic motivation. Especially for young students, tangible rewards seem to have a stronger detrimental effect on intrinsic motivation than such rewards can have on college students (Deci et al., 1999). However, Levitt et al. (2012) found no strong evidence of crowding out in their study of 6,500 elementary and high school students. Bettinger and Slonim (2007) also found no evidence that a test performance incentive program is detrimental to elementary school students' intrinsic motivation.

Extrinsic motivation can come in many different forms. While a majority of research has focused on the implementation of financial rewards, there is evidence that the effect of non-monetary rewards can be considerable. Kosfeld and Neckermann (2010) discovered that non-material rewards such as awards or trophies can have significant motivational power, stating that awards yield non-material benefits in the form of social recognition, status and improved self-esteem. Besley and Ghatak (2008) found that status incentives increase effort while reducing the optimal level of monetary incentives. Levitt et al. (2012) found considerable support for the effect of both financial and non-financial rewards on

short-term student effort and performance.

2.2 Research on incentives

Many traditional methods in school use anxiety incited by the threat of bad grades as a form of negative motivation to achieve higher student performance. However, research has demonstrated that positive motivation (as opposed to threat of punishment) is more effective for encouraging students to learn. In fact, the anxiety produced by negative motivation often hinders performance of complex tasks (Moen and Doyle Jr, 1978). The Yerkes-Dodson law brings further evidence to this claim, as research has found that there is an empirical relationship between arousal and performance (Yerkes and Dodson, 1908). Performance on simple tasks increases with stress or arousal, which is seen as an energizing effect, but diminishes on difficult tasks, which is caused by the negative effects of stress or arousal on cognitive abilities (Diamond et al., 2007).

Ranking students by their performance is another important and effective tool that can be used as a source of motivation, as rank in itself functions as a major motivator. Tran and Zeckhauser (2012) confirmed that the desire for rank in wealth or status has a measurable impact on behavior. Rank can be used by students to impress friends and family and to earn respect and admiration, which can be seen as tangible benefits. Humans may also be directly psychologically rewarded by higher rank without the need for any tangible benefits. Students who receive rank publicly outperform those who receive their rank privately, but even when ranking information cannot be reliably communicated, there is still an effect on performance (Tran and Zeckhauser, 2012). Rank as a motivator in school can often be seen in the form of norm-referenced grading, where students are assigned grades relative to the performance of other students.

Levitt et al. (2012) verified that in the absence of immediate incentives, students tend to exert low effort on standardized tests. These standardized tests are often of little importance to the students, so called low-stake tests, but have important consequences for the teachers (e.g., in the form of allocation of resources), and are as such high-stake

tests for the teachers. Attali et al. (2011) showed that males exhibit a larger difference in performance between low and high-stake tests than females do, and also found support for differences with respect to socioeconomic factors such as student background. Students with a higher socioeconomic status showed larger differences in performance.

Levitt et al. (2012) found that the effects of introducing extrinsic motivators were larger for boys than for girls, i.e. that boys are more responsive to short-term incentives than girls. This suggests that girls may be more intrinsically motivated, and therefore also be at a higher risk of crowding out. In their study, Levitt et al. (2012) examined various types of motivators, including financial and non-financial incentives, immediate rewards and rewards handed out with a delay, as well as rewards framed as losses. They found that incentives framed as losses (giving the reward before the test and taking it back if a specified goal is not met) had a stronger effect than other incentives, and that non-financial incentives were effective on elementary school students but had little effect on older students. They also found that delayed rewards had no motivational power. This has important implications for the way the educational system is currently set up, with almost all feedback coming with a delay.

Wise and DeMars (2005) discuss a number of potential assessment practices for managing the problems posed by low student motivation leading to lower test performance. The issue of students not giving full effort on a test is of critical importance to assessment practitioners, as the results will tend to underestimate the students' true skill levels. Wise and DeMars' results show that motivation is an essential factor in test performance and that higher motivation is associated with higher test scores. Motivation is therefore an important factor in eliciting test results that accurately reflect a student's true knowledge and skill level.

2.3 Our approach

With this background of previous research, we wish to contribute by evaluating different, cost-effective methods of incentivizing younger students. To the best of our knowledge,

there has been no study focusing purely on non-financial incentives in elementary school. There is also a prominent gap in Swedish research when it comes to student test-taking motivation, and as Swedish students have been shown to underperform relative to their neighbors, we recognize the importance of investigating this further. Our interest lies in evaluating the effects of different, non-financial incentives on test-taking performance. We want to examine how test scores of incentivized students differ from those of unincentivized students, on average. We are also interested in analyzing the effects with respect to gender to see if boys and girls, on average, react differently to the same incentives.

3 Method

3.1 Experimental design

To investigate the motivational power of non-financial incentives on elementary school students, we conducted a randomized experiment. Students in the experiment were assigned to one out of five groups; either to an unincentivized control group or to an incentivized treatment group. By randomly allocating the control and four different treatments within each class, we are able to examine the effect of one incentivizing factor at a time, and minimize the impact of endogenous variables such as school and class-specific factors. Students were offered no choice in whether to participate or not, and therefore we eliminate the potential sample bias that could arise with self-selection and voluntary participation. The randomization process means that we obtain groups that are statistically equivalent to each other, and we can thereby simply compare the difference between the means of the treatment groups and the control group. This provides an internally valid estimate of the causal effect of treatment.

Although the experimental approach circumvents the problem of selection bias and offers the virtue of internal validity, it does bring some potential issues regarding environmental dependence and replicability. We can in no way guarantee that our results would be valid if the experiment were to be repeated in a different context. As our experimental design poses a threat to external validity, its results should best be interpreted as what *can* happen but not necessary what *will* happen in an external environment where other variables are free to operate without being tightly controlled.

3.2 Implementation

The experiment was carried out on 437 sixth grade students in a total of 22 classes in ten elementary schools in the Stockholm municipality. Each school in the City of Stockholm's directory of elementary schools was given a number, and with the aid of an online randomizer, we were able to randomize the selection of schools to contact. Teachers were contacted by telephone and asked to participate in our experiment. The sessions were usually carried out one to two weeks after the phone call was made. Out of seventeen contacted schools, ten accepted. The only expressed reason for not participating was the heavy student workload, however we assess that this would have been more or less equivalent irrespective of school. We do not suspect that the teachers' choice of participating should have led to a biased sample, as all participating schools showed a wide variety of school-specific characteristics. Furthermore, teachers received limited information on the specifics of the experiment. We therefore see the choice of accepting or declining as random, or at the very least not correlated with the nature of the experiment. Schools accepting the study were found to be represented both on the high-performing and low-performing ends of the spectrum, and they were diverse with respect to geographic location and socioeconomic factors.

The experiment took place in April 2013 during scheduled lecture hours and consisted of a standardized mathematical test containing four tasks giving a maximum of 22 points in total. The tasks matched the level of difficulty of tasks in the Swedish national exams. Furthermore, they were designed with support from educated elementary school teachers not present at any of the schools where the experiment was conducted. They were formulated in such a way as to allow efficient and impartial grading.

All sessions were introduced and held entirely by ourselves but in the presence of the teacher. This ensured that the experiment was perceived as formal, encouraging students to take the test seriously. It was presented in a formal way so as to establish commitment to the task. Special care was taken to ensure that the experiment was presented in an equal, or at least in a very similar, way for all classes and schools. The aim was for the

experiment to be perceived equally by all participating students. However, this is nothing we are able to confirm. Some students may have viewed the test as more important than others and as such applied more effort, but overall, no systematic deviations should exist between the groups.

In each class, we randomly assigned students to control and treatment groups by handing out tests with differing information concerning the assessment of their performance. We did this in a randomized fashion. To prevent any kind of preparation, teachers had received limited information regarding the formalities of the test. Just before the test started, we stressed the importance of solving the test individually, in silence, and of carefully reading all the information provided. Students were given ten minutes to solve the test. Questions regarding how to think about the problems were responded to with the same, limited information. We asked students to remain seated with the test in front of them until the ten minutes had passed, thus avoiding any potential benefit that could arise from finishing the test early.

When the time had passed, we immediately collected and corrected the tests. Subsequent to the assessment, we returned to the classroom and qualifying students received their rewards. The class was also told the purpose of the test and our experiment, and students were able to take a look at their test score.

3.3 Choice of treatment variables

The treatment variables we chose to evaluate reflect our interest in analyzing cost-effective, extrinsic incentives in the educational setting. All chosen treatments are non-financial, and we also analyze the effect of norm and criterion-based grading.

All students received the same test but at the top of the test, students received different information depending on their group assignment. Subjects in the control group received no information regarding the assessment of the test. The only information they received was the total amount of points obtainable. Subjects in treatment group 1 received the same test but were informed that their performance would be graded on the scale

Table 3.1: Control group and treatment groups with corresponding incentives

Group	Grade A-F	Grade A	Diploma	Prize	Top 3
Control					
Treatment 1	×				
Treatment 2		×			×
Treatment 3			×		
Treatment 4				×	×

A-F. They were also given the scale of points corresponding to each grade. Subjects in treatment group 2 were informed that the top three performing students in the class would receive the grade A. Subjects in treatment group 3 were informed that obtaining a score of 18 or above would result in receiving a diploma. Subjects in treatment group 4 were informed that the top three performing students would receive a prize. Table 3.1 displays a summary of the treatments carried out in the different groups.

The literal information given to the students can be seen in its original Swedish form with corresponding translations in Table 3.2. The complete test with its tasks can be found in Appendix.

The choices of treatments for group 1 and group 3 relate to the theory behind criterion-referenced assessment, which involves determining a grade by comparing a student's achievement with clearly stated criteria for learning outcomes for a particular performance level. The groups differ with respect to assessment as group 1 receives grading and group 3 receives a diploma. We are interested in comparing the effects of grades as incentives to that of a symbolic reward such as a diploma. In group 3, a test score of 18 or above is the required level of achievement in order to receive a diploma. 18 points is chosen as the cut-off threshold as this is determined to be an obtainable amount through the application of higher effort. The usage of a higher cut-off than 18 was deemed to be perceived as less likely obtainable and have a demotivating effect. We compare this type of assessment to that of norm-referenced grading, in which a student's grading is based on their relative ranking within a particular group of students. Norm-referenced grading involves fitting a ranked list of students' scoring to a pre-determined distribution for rewarding grades. This type of grading can be seen in treatment groups 2 and 4, where it is stated that only the top three performing students will be rewarded. Norm-referenced

Table 3.2: Information regarding test assessment

Control group
På detta test kan du få totalt 22 poäng.
On this test you can obtain a total of 22 points.
Treatment group 1
På detta test kan du få totalt 22 poäng.
On this test you can obtain a total of 22 points.
Om du får 21-22 poäng får du ett A.
If you obtain 21-22 points you will receive an A.
Om du får 18-20 poäng får du ett B.
If you obtain 18-20 points you will receive a B.
Om du får 15-17 poäng får du ett C.
If you obtain 15-17 points you will receive a C.
Om du får 12-14 poäng får du ett D.
If you obtain 12-14 points you will receive a D.
Om du får 10-11 poäng får du ett E.
If you obtain 10-11 points you will receive an E.
Om du får 0-9 poäng får du ett F.
If you obtain 0-9 points you will receive an F
Treatment group 2
På detta test kan du få totalt 22 poäng.
On this test you can obtain a total of 22 points.
Om du är bland de tre med högst antal poäng i klassen får du ett A.
If you are among the three with the highest score in the class you will receive an A.
Treatment group 3
På detta test kan du få totalt 22 poäng.
On this test you can obtain a total of 22 points.
Om du får 18 poäng eller mer får du ett diplom.
If you obtain 18 points or more you will receive a diploma.
Treatment group 4
På detta test kan du få totalt 22 poäng.
On this test you can obtain a total of 22 points.
Om du är bland de tre med högst antal poäng i klassen får du ett pris.
If you are among the three with the highest score in the class you will receive a prize.

grading can be used as a motivation tool as it speaks to the students' desire to be ranked highly. It has been shown that students who receive rank outperform those who do not (Tran and Zeckhauser, 2012).

We chose these treatments due to our interest in comparing the differing effects of the traditional incentive of grades without ranking to those of alternative incentives such as diplomas, prizes and grades with ranking. The choice of diploma as a reward reflects an interest in analyzing non-material status rewards, as Kosfeld and Neckermann (2010)

found that these types of rewards can have a significant impact and motivational power. A prize is the only material reward and is therefore used to compare materialistic incentives to non-materialistic incentives.

3.4 Econometric model

Our aim is to estimate if different, non-financial incentives applied in the test-situation have causal effects on student performance, where student performance is measured as obtained test score. We now introduce the methodology which allows us to estimate the Average Treatment Effect (ATE) for the different proposed treatments. The ATE of a treatment (incentive) T on the outcome (test score) y for the experimental unit (student) i, can be defined by comparing the outcomes that would have occurred under each of the different treatment possibilities. Using the potential outcome notation popularized by Rubin (1974), let y_{ji} be the outcome for each student i under a treatment j and y_{0i} be the outcome under control. For each student i in treatment j, we observe T_{ji} and y_{ji}, where y_{ji} is defined as

$$y_{ji} = y_{0i} + (y_{ji} - y_{0i})T_{ji}. \tag{3.1}$$

The value $y_{ji} - y_{0i}$ is the treatment effect for treatment j in student i and T_{ji} takes value 1 for treatment j, and 0 otherwise. If one could observe y_{0i} and y_{ji} for each individual, one could estimate the ATE by analyzing the average value of $y_{ji} - y_{0i}$ for all students in each treatment. However, we are able to observe either y_{0i} or y_{ji}, never both, since each student is subject to exactly one of the four treatments or control. To estimate the ATE, we resort to ordinary least squares (OLS) estimation. Thus, we can rewrite Eq. (3.1) as:

$$Test\_score_{ji} = \alpha_0 + \delta_j T_{ji} + \epsilon_{ji}, \;\; j = 1, ..., 4, \tag{3.2}$$

where α_0 is the control group average test score, δ_j the average causal effect of treatment j and ϵ_{ji} is the residual. We then estimate the regression given by Eq. (3.2) to obtain the average causal effect of the different treatments. The complete randomization of assigning

students to control group and treatment groups assures that T_{ji} is uncorrelated with ϵ_{ji}. This yields an unbiased estimator for the average causal effect of treatment. To ensure robust results, we control for gender, class size and school-specific factors. These are factors that may causally affect the outcome variable. We include the binary variable $female$ to control for gender differences and the variable $class\_size$ to control for class size. We let X_s denote a vector of covariates measured at the school-level, representing GPA, percentage of foreign-born students, percentage of Swedish-born students with foreign-born parents and average parent education level. The average causal effect of treatment j, δ_j, is estimated from the following equation:

$$Test\_score_{ji} = \alpha_0 + \delta_j T_{ji} + \gamma_j \cdot female + \phi_j \cdot class\_size + \beta_j X_s + \epsilon_{ji}, \ j = 1, ..., 4. \quad (3.3)$$

We can subsequently employ hypothesis testing where we test the following null and alternative hypothesis.

$$\begin{aligned} H_0 &: \delta_j = 0 \\ H_1 &: \delta_j \neq 0 \end{aligned} \quad (3.4)$$

The above represents that the null hypothesis of no average causal effect of treatment is tested against the alternative that there is a non-zero average effect of treatment. If we can reject the null hypothesis there are grounds to believe that the treatment has a measurable effect on the outcome variable test score.

4 Data

The experiment was carried out on a total of 437 students, but due to implementation difficulties in one of the schools, we have chosen to exclude the results obtained in that particular school from our dataset. The experiment was unsupervised in one of three classes and students had been given prior, misleading information about the test and its implications before our arrival at the school. Including these observations leads to biased estimates. However, we have confirmed that it does not change their significance. We have chosen not to include these observations due to our concern about the way the data was gathered. Our cross-sectional dataset therefore consists of a total of 378 observations.

Of our 378 observations, 170 observations are boys and 208 are girls. The gender distribution across the groups can be seen in Table A.1 in Appendix. The number of students in each group spans from 73 to 77. The scores obtained range from the minimum of 0 points to the maximum of 22 points and the test score distribution is negatively skewed, which means that the students perceived the test as relatively easy. See Fig. A.1 in Appendix.

As a result of randomization within each class, we obtained a balanced number of students across treatment and control groups. This randomization also implies that the groups are balanced with regard to other factors such as gender, socioeconomic background and school and class-specific factors. We have chosen to include a set of control variables in our dataset to increase the precision of our results and to certify that our findings are robust. The first control variable is female which indicates gender. The second control variable is average GPA at graduation for each school, which we are using as a proxy variable for school quality. The third control variable is class size, as differences in class size could have an indirect effect on student learning and skill level through teacher-student time and attention. The fourth and fifth control variables relate to socioeconomic background of the students in the school as a whole, with one being the percentage of

foreign-born students present at the school, the other being the percentage of students born in Sweden with both parents being foreign-born. The sixth and final control variable that is included is a measure of parent education level. With the exception of gender and class size, all data on the control variables were obtained at the school level from the analysis tool Skolverkets Arbetsverktyg för Lokala SambandsAnalyser (SALSA), which fills the purpose of highlighting factors that could impact average grades, by school. The tool is administered by The Swedish National Agency for Education and is based on statistics gathered from Statistics Sweden's school register. Data was available only for schools with grades 1-9, and out of the schools in our sample, two did not fulfil this criteria.

To ensure that our dataset is balanced, we have conducted t-tests for each control variable. Table A.2 in Appendix displays t-statistics for the test of equal means between control group and treatment groups. Since no value is of statistical significance, we can conclude that treatment groups are statistically equal to the control group. This is true for all control variables. This is an important result as it implies that we will obtain unbiased estimates of the causal effect of treatment.

5 Results

Table 5.1 reports means for the control group with respect to chosen variables, and the differences in means between the control group and each treatment group. The only differences in means that are of statistical significance are those for test scores for treatment groups 2-4. When comparing means of test scores in treatment groups against the control group, we see that all treatment groups show a positive difference in mean test score. This indicates that students in the incentivized treatment groups performed better than students in the unincentivized control group, on average. Included control variables are statistically equal in means, indicating that our data shows no sign of suffering from a sample bias.

Table A.3 in Appendix reports the group mean test scores obtained in each of the control and treatment groups. The average test score obtained in the experiment was 14.91. We observe the highest average points in treatment group 4 (prize), and the lowest average points in the control group; 15.71 compared to 13.50. This implies that the treatment with the greatest impact in our experiment was being incentivized with a prize, and that receiving no information regarding assessment or other incentives led to the lowest performance. The average score in treatment group 1 (grade A-F) was 14.55, in treatment group 2 (grade A) 15.55, and in treatment group 3 (diploma) 15.26.

When analyzing test scores with respect to gender, we observe a mean score of 14.52 for boys and 15.23 for girls. This shows that girls on average performed better than boys in the experiment. We observe the highest average test score for boys in treatment group 2 (grade A), with a mean score of 15.67, and the highest average test score for girls in treatment group 4 (prize), with a mean score of 16.05. The descending order of average test scores for boys is the following: group 2 (grade A) 15.67, group 4 (prize) 15.25, group 3 (diploma) 14.71, group 1 (grade A-F) 13.74 and control group 13.49. The descending order of average test scores for girls is the following: group 4 (prize) 16.05,

Table 5.1: Differences in means

	Control mean	T1-C	T2-C	T3-C	T4-C
Test Score (points)	13.500	1.053	2.048**	1.760**	2.211***
	(0.599)	(0.836)	(0.807)	(0.807)	(0.799)
Test Score (standardized)	-0.287	0.214	0.417**	0.358**	0.450***
	(0.122)	(0.170)	(0.164)	(0.164)	(0.163)
Female	0.487	0.053	0.061	0.111	0.092
	(0.057)	(0.082)	(0.082)	(0.081)	(0.081)
Class Size	20.789	0.211	-0.132	0.353	0.316
	(0.564)	(0.789)	(0.800)	(0.798)	(0.818)
GPA	32.349	0.120	-1.383	-1.318	0.909
	(4.250)	(6.020)	(6.103)	(6.037)	(6.154)
Foreign-Born	0.147	0.005	0.007	0.012	0.006
	(0.019)	(0.028)	(0.028)	(0.028)	(0.028)
Foreign Parents	0.149	0.000	0.006	0.002	-0.005
	(0.014)	(0.020)	(0.021)	(0.020)	(0.020)
Parent Education	2.357	-0.001	-0.013	-0.015	0.000
	(0.032)	(0.046)	(0.046)	(0.046)	(0.046)

Note: the first column presents the mean for control group of the variable indicated in each row. Columns T1-C to T4-C represent the difference in mean between treatment groups and control group. Robust standard errors are displayed in parentheses. Asterisks next to coefficients indicate a significant difference of means, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

group 3 (diploma) 15.61, group 2 (grade A) 15.45, group 1 (grade A-F) 15.24 and control group 13.51. The differences in average test score between the different groups for all students and with respect to gender can be seen in Figs. A.2 and Table A.3 in Appendix. For girls, the average test scores for all treatment groups are significantly higher than the average test score for the control group. For boys, the respective test scores are more spread.

We run regressions (3.2) and (3.3). OLS estimates for raw test scores are reported in Tables 5.2 and 5.3. OLS estimates for standardized test scores are reported in Tables A.4 and A.5 in Appendix. In what follows, we analyze regressions on raw test score. We find this more informative considering the skewness of the distribution. Tables 5.2 and 5.3 show average treatment effects in comparison to the control group, and we separate the effects between the group as a whole from the effects for boys and for girls separately. With the exception of treatment group 1, all the treatment groups are significant at the

5% level or less, and the constant is significant at the 1% level. The treatment group with the largest coefficient is treatment group 4 (prize), with δ_4 of 2.21 and a significance at the 1% level. This indicates that receiving the information that the performance on the test may lead to being rewarded with a prize is associated with a 2.21-point higher score than the control group on average. The second largest effect of treatment can be seen in treatment group 2 (grade A), where assignment to this group is associated with a 2.05-point higher score on average, compared to the control group. This result is significant at the 5% level. A smaller effect can be seen in treatment group 3 (diploma), where the average difference in score is 1.76 points compared to the control group. This result is significant at the 5% level. Being assigned to treatment group 1 (in which students receive information that they will be graded on the scale A-F) shows no significant difference in average test score.

When expanding the regression to include the control variables gender, class size, average school GPA, percentage of foreign-born students, percentage of Swedish-born students with both parents being foreign-born, and parent education level, our results remain robust. Previously significant coefficients remain significant on at least a level of 5%.

Examining boys separately yields significant results only for treatment group 2 (grade A), with a significance at the 10% level. The constant, representing the control group, remains significant at the 1% level, with a value of 13.49. These results indicate that for boys, the only effective treatment is receiving the information that the top three performing students will be given the grade A. Comparing this result to those obtained when looking only at test scores for girls, we find interesting differences. For girls, all treatment groups with the exception of group 1 (grade A-F) are significant at the 10% level or less. The treatment showing the highest effect on performance is group 4 (prize), with δ_4 of 2.53 and a significance level of 5%. This indicates that girls being incentivized with the information that they may receive a prize have a 2.53-point higher score than the control group on average. For girls, the assignment to treatment group 3 (diploma) shows the second largest difference in test scores as compared to the control group, more

Table 5.2: Impact of incentives on test scores for all students

	Without Controls	With Controls
T1	1.053	1.111
	(0.836)	(0.852)
T2	2.048**	2.596***
	(0.807)	(0.829)
T3	1.760**	1.968**
	(0.807)	(0.836)
T4	2.211***	2.711***
	(0.799)	(0.836)
Female		0.374
		(0.548)
GPA		-0.000774
		(0.0375)
Class Size		0.158
		(0.0961)
Foreign-Born		7.962**
		(3.647)
Foreign Parents		4.988
		(3.331)
Parent Education		9.539*
		(5.228)
Constant	13.50***	-14.89**
	(0.599)	(6.877)
Observations	378	313
R-squared	0.027	0.176

Note: The table reports OLS estimates. Robust standard errors are displayed in parentheses. Asterisks next to coefficients indicate a significant difference of means, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

specifically this group receives a 2.10-point higher test score on average. Assignment to treatment group 2 (grade A), shows an average 1.94-point higher test score compared to the control group. These result are significant at the 10% level.

When including the aforementioned control variables, we find that for boys, the coefficient for treatment group 2 (grade A) remains significant, and that the coefficient for treatment group 4 (prize) becomes significant at a 5% level. This treatment group did not show significant results without the inclusion of the control variables. This implies that we increased the precision of our results. For girls, the inclusion of the control variables does not lead to any changes in significance of the treatment variable coefficients, and they all remain significant at a minimum of the 5% level.

Table 5.3: Impact of incentives on test scores by gender

	Boys		Girls	
	Without Controls	With Controls	Without Controls	With Controls
T1	0.256	0.594	1.730	1.744
	(1.237)	(1.294)	(1.129)	(1.135)
T2	2.179*	3.124**	1.936*	2.578**
	(1.288)	(1.445)	(1.024)	(0.999)
T3	1.255	1.312	2.095*	2.710**
	(1.226)	(1.273)	(1.071)	(1.114)
T4	1.763	2.495**	2.532**	3.208***
	(1.177)	(1.227)	(1.084)	(1.144)
GPA		0.0161		-0.0331
		(0.0717)		(0.0460)
Class Size		0.0222		0.304**
		(0.142)		(0.140)
Foreign-Born		11.84**		3.291
		(5.709)		(4.790)
Foreign Parents		5.967		5.175
		(5.346)		(4.528)
Parent Education		11.81		8.971*
		(11.68)		(5.284)
Constant	13.49***	-21.72	13.51***	-8.573
	(0.897)	(14.21)	(0.800)	(7.831)
Observations	170	141	208	172
R-squared	0.029	0.158	0.031	0.207

Note: The table reports OLS estimates. Robust standard errors are displayed in parentheses. Asterisks next to coefficients indicate a significant difference of means, where $(***,\ p < 0.01)$, $(**,\ p < 0.05)$ and $(*,\ p < 0.1)$.

We graphically illustrate the test score distributions for each group, both for the students as a whole and for boys and girls respectively. This can be seen in Figs. A.3 to A.7 in Appendix. We note that with treatments in place, the test score distributions are negatively skewed towards higher points. For boys the largest negative skew is present for treatment group 2 (grade A), and for girls the largest skew is seen for treatment group 4 (prize). This is in accordance with our results above.

Figs. A.8 to A.10 in Appendix give further nuance to the test score characteristics. We note that treatment group 2 has a high number of students contained within a relatively small segment of the sample. The median test score is among the highest median test scores of all groups, together with that of treatment group 3. We can see from the plot

that the negative skewness of test scores is the most prominent for these two groups. We also note the relatively high number of outliers in treatment group 2. Box plots displaying test scores for boys and girls separately show similar characteristics. Note that the median test score for girls is above the general average, while the median test score for boys is below the general average.

6 Discussion

The results from our experiment show that extrinsic incentives have a motivational effect on student performance in a test situation. To rank students by distinguishing the top three performers has particularly large motivational power. However, the motivational power of evaluated incentives differ with respect to gender. Boys appear to apply the most effort and perform the best when they are incentivized with receiving the information that they can obtain grade A if they are among the top three performing students in the class. Girls, on the other hand, appear to be strongest motivated by being informed that they can receive a prize if they are among the top three performing students in the class. The only incentive that appears to have no significant effect on student effort and performance is the incentive of being graded on the scale A-F. This is a striking result considering the way the educational system is currently constructed. To grade with criteria, without a corresponding distribution, is the most commonly used method of student assessment in Swedish elementary schools. Our results suggest that this method may not be optimal for extracting high student effort in test situations.

As previous research implies, norm-based grading has a much more significant impact on student effort and performance than criterion-based grading. This is confirmed by our results. However, it is interesting to observe the differences in the level of impact between boys and girls that norm-based grading has. Boys are surprisingly prone to responding to being ranked with respect to grades whereas girls react surprisingly strongly to the incentive of receiving a prize. Intuitively, we would have expected the opposite to be true, as receiving a prize offers a more public form of competition, and boys have been shown to exhibit more competitive behavior (Croson and Gneezy, 2009; Niederle and Vesterlund, 2010). It is interesting that we found the opposite to be true, but we cannot be certain of how the incentives were perceived by the students, and thus we cannot conclude that prize was the most worthy of competition. Therefore, boys may still have displayed the

most competitive behavior but have valued grade A higher than prize. As prize is the only materialistic reward tested, girls' strong reaction to this incentive may indicate that they are more materialistic than boys.

Our results demonstrate that incentives have a much larger general effect on girls than on boys. For girls, we can see a considerable impact on performance as soon as an incentive is in place, but for boys this impact is not as prevalent. Boys have a greater spread in, and a higher standard deviations of, test scores as compared to girls.

There is also a general gender difference in which girls outperform boys with an average of 0.71 points. This tells us either that girls apply more effort in the test situation, or that they are simply better at solving the mathematical tasks presented in the experiment. Either way, it accords with our experience from hosting the experiments in the schools. Our impression was that girls took the test more seriously than boys. In general, they were also more curious to know their test score.

7 Conclusions

The educational system is built upon a continuous usage of tests as a measure of student performance and the results of these tests often lay the foundation for much of the educational debate. For tests to be a useful measure of student knowledge and performance, they need to accurately reflect students' skill levels. This can only be the case if students are motivated to perform well when taking tests, and if they exert enough effort in the test situation.

Our study on student motivation and test-taking behavior at the elementary school level in Sweden shows that there is cause to question the current incentives being used to motivate students. Through a randomized experiment, we have discovered that the system of grading students according to a pre-specified set of criteria is inefficient. This is the most widely-used assessment method for Swedish elementary school students today, and therefore, our results have important implications.

We believe that there are other avenues of research on this topic. By further researching other motivational incentives, and by increasing the sample size, one could find even more efficient methods for maximizing student performance. From the methods that we studied, we found those involving norm-referenced assessment to be the most effective. Rank-based grading is often prevalent at later educational stages, but our results indicate that an earlier implementation could have positive effects on student performance. Our experiment has also highlighted gender differences with regard to the efficiency of the incentives implemented. In order to maximize performance levels, the knowledge derived from our study could be utilized.

Bibliography

Y. Attali, Z. Neeman, and A. Schlosser. Rise to the challenge or not give a damn: differential performance in high vs. low stakes tests. *IZA Discussion Paper No. 5693*, 2011.

T. Besley and M. Ghatak. Status incentives. *American Economic Review*, 98(2):206–211, 2008.

E. Bettinger and R. Slonim. Patience among children. *Journal of Public Economics*, 91 (1):343–363, 2007.

R. Croson and U. Gneezy. Gender differences in preferences. *Journal of Economic Literature*, pages 448–474, 2009.

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

D. M. Diamond, A. M. Campbell, C. R. Park, J. Halonen, and P. R. Zoladz. The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the yerkes-dodson law. *Neural Plasticity*, 2007.

G. Eisenkopf. Paying for better test scores. *Education Economics*, 19(4):329–339, 2011.

Ekonomifakta. Resultat PISA - internationellt, 2009. URL http://www.ekonomifakta.se/sv/Fakta/Utbildning-och-forskning/Provresultat/. (Accessed 15 May 2013).

R. G. Fryer. Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, 126(4):1755–1798, 2011.

M. Kosfeld and S. Neckermann. Getting more work for nothing? symbolic awards and worker performance. *IZA Discussion Paper No. 5040*, 2010.

S. D. Levitt, J. A. List, S. Neckermann, and S. Sadoff. The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. Technical report, National Bureau of Economic Research, 2012.

W. Margaret and A. Ray. *PISA Programme for International Student Assessment (PISA) PISA 2000 Technical Report: PISA 2000 Technical Report*. OECD Publishing, 2003.

M. R. Moen and K. O. Doyle Jr. Measures of academic motivation: A conceptual review. *Research in Higher Education*, 8(1):1–23, 1978.

M. Niederle and L. Vesterlund. Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives*, 24(2):129–144, 2010.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

A. Tran and R. Zeckhauser. Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 2012.

S. L. Wise and C. E. DeMars. Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1):1–17, 2005.

R. M. Yerkes and J. D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482, 1908.

A Appendix

Table A.1: Student distribution across groups

Group	N	Boys	Girls
C	76	39	37
T1	76	35	41
T2	73	33	40
T3	77	31	46
T4	76	32	44
total	378	170	208

Table A.2: t-statistics testing for equal means

	T1	T2	T3	T4
Female	-0.6457	-0.7425	-1.3720	-1.1354
	0.5194	0.4590	0.1721	0.2580
Class Size	-0.2669	0.1658	-0.4431	-0.3863
	0.7899	0.8685	0.6584	0.6998
GPA	-0.0199	0.2266	0.2183	-0.1477
	0.9842	0.8211	0.8275	0.8828
Foreign-Born	-0.1648	-0.2339	-0.4436	-0.1978
	0.8694	0.8155	0.6581	0.8435
Foreign Parents	0.0149	-0.2811	-0.1150	0.2219
	0.9882	0.7791	0.9086	0.8247
Parent Education	0.0127	0.2899	0.3323	0.0010
	0.9899	0.7724	0.7402	0.9992

Note: Values in columns represent t-statistics from hypothesis testing that mean value in control group equals mean value in treatment group. Values below t-statistics represent p-values. Rows represent different control variable averages tested. Asterisks next to t-values indicate a significant difference, where $(***,\ p < 0.01)$, $(**,\ p < 0.05)$ and $(*,\ p < 0.1)$.

Table A.3: Average test scores across groups

	Control	T1	T2	T3	T4
N	76	76	73	77	76
Average test score (All)	13.50	14.55	15.55	15.26	15.71
	(0.599)	(0.584)	(0.542)	(0.541)	(0.530)
Average test score (Boys)	13.48	13.74	15.67	14.74	15.25
	(0.895)	(0.851)	(0.924)	(0.837)	(0.763)
Average test score (Girls)	13.51	15.24	15.45	15.61	16.05
	(0.801)	(0.796)	(0.639)	(0.711)	(0.731)

Note: Robust standard errors are displayed in parentheses.

Table A.4: Impact of incentives on test scores for all students, standardized

	Without Controls	With Controls
T1	0.0860	0.0908
	(0.0684)	(0.0696)
T2	0.165**	0.209***
	(0.0650)	(0.0667)
T3	0.145**	0.162**
	(0.0663)	(0.0687)
T4	0.181***	0.222***
	(0.0653)	(0.0683)
Female		0.0379
		(0.0556)
GPA		-0.00537
		(0.260)
Class Size		0.158
		(0.0960)
Foreign-Born		0.255**
		(0.117)
Foreign Parents		0.116
		(0.0773)
Parent Education		0.498*
		(0.273)
Constant	-5.93e-09	-0.0820
	(0.0510)	(0.0559)
Observations	378	313
R-squared	0.027	0.176

Note: The table reports OLS estimates. Variables have been rescaled to have a mean of zero and standard deviation of one. Robust standard errors are displayed in parentheses. Asterisks next to coefficients indicate a significant difference of means, where (***, $p < 0.01$), (**, $p < 0.05$) and (*, $p < 0.1$).

Table A.5: Impact of incentives on test scores by gender, standardized

	Boys		Girls	
	Without Controls	With Controls	Without Controls	With Controls
T1	0.0209	0.0485	0.141	0.143
	(0.101)	(0.106)	(0.0922)	(0.0927)
T2	0.175*	0.251**	0.156*	0.208**
	(0.104)	(0.116)	(0.0824)	(0.0804)
T3	0.103	0.108	0.172*	0.223**
	(0.101)	(0.105)	(0.0880)	(0.0915)
T4	0.144	0.204**	0.207**	0.262***
	(0.0962)	(0.100)	(0.0886)	(0.0935)
GPA		0.111		-0.229
		(0.496)		(0.319)
Class Size		0.0222		0.303**
		(0.142)		(0.140)
Foreign-Born		0.380**		0.105
		(0.183)		(0.154)
Foreign Parents		0.138		0.120
		(0.124)		(0.105)
Parent Education		0.617		0.469*
		(0.610)		(0.276)
Constant	-0.0693	-0.0989**	0.0532	-0.0780
	(0.0779)	(0.0853)	(0.0673)	(0.0822)
Observations	170	141	208	172
R-squared	0.029	0.158	0.031	0.207

Note: The table reports OLS estimates. Variables have been rescaled to have a mean of zero and standard deviation of one. Robust standard errors are displayed in parentheses. Asterisks next to coefficients indicate a significant difference of means, where $(***,\ p < 0.01)$, $(**,\ p < 0.05)$ and $(*,\ p < 0.1)$.

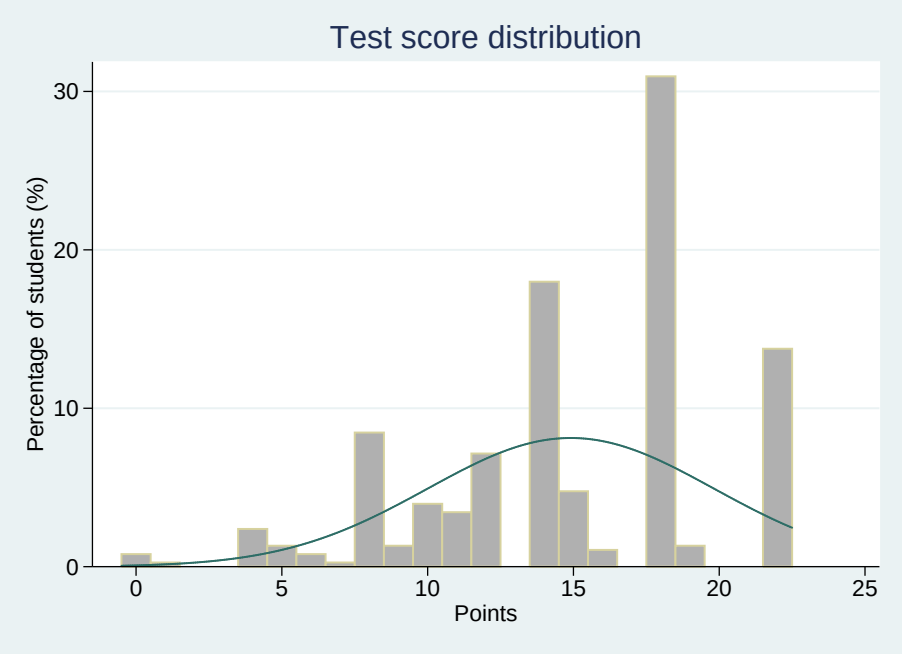
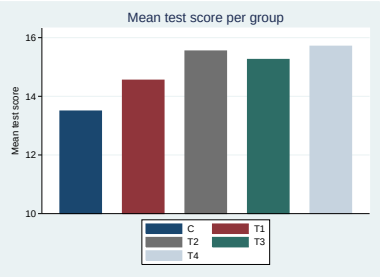
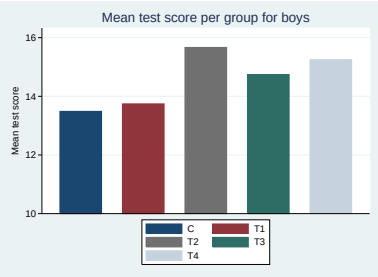


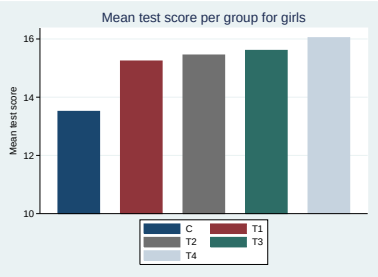
Figure A.1: Test score distribution



(a)

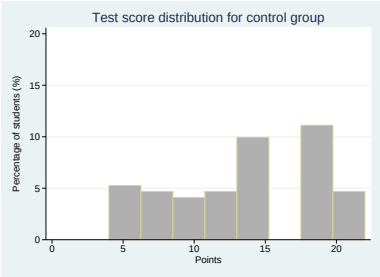


(b)

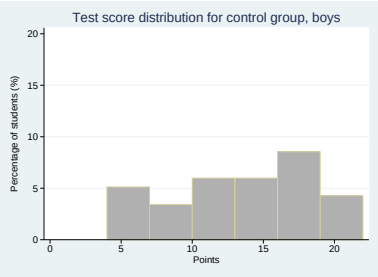


(c)

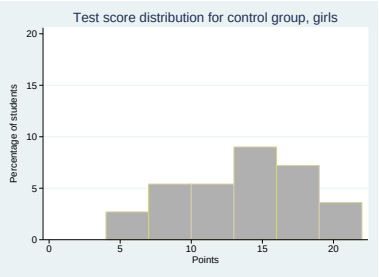
Figure A.2: Mean test score per group for all students and by gender



(a)

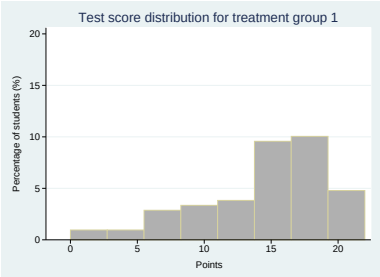


(b)

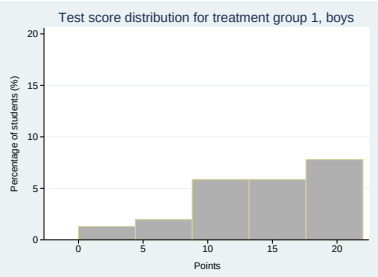


(c)

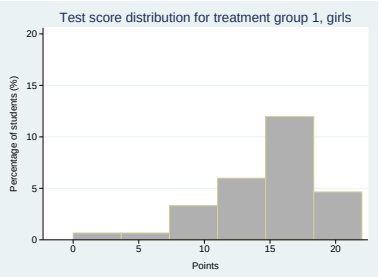
Figure A.3: Control group test score distributions for all students and by gender



(a)

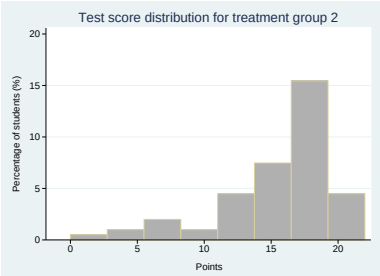


(b)

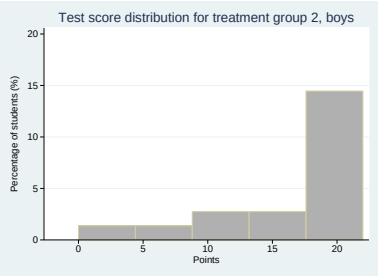


(c)

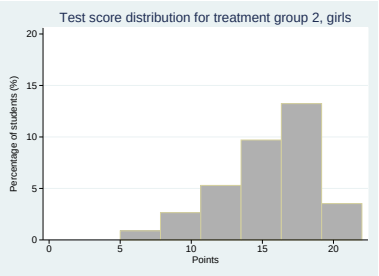
Figure A.4: Treatment group 1 test score distributions for all students and by gender



(a)

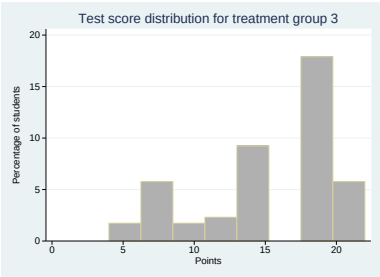


(b)

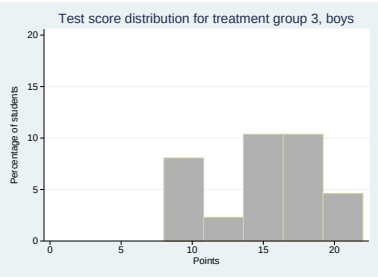


(c)

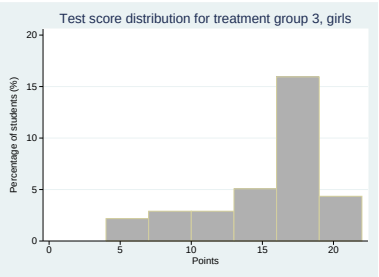
Figure A.5: Treatment group 2 test score distributions for all students and by gender



(a)

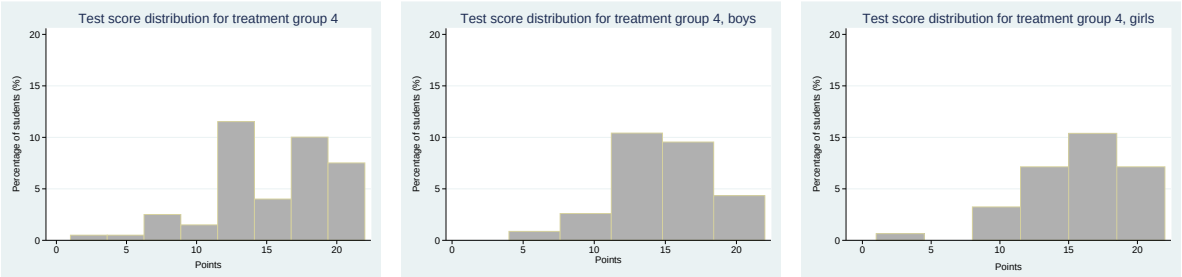


(b)



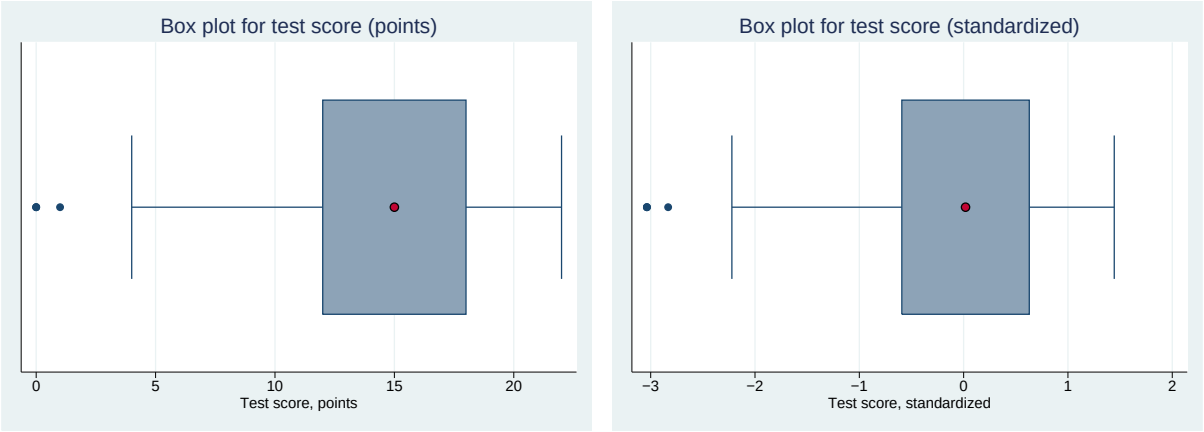
(c)

Figure A.6: Treatment group 3 test score distributions for all students and by gender



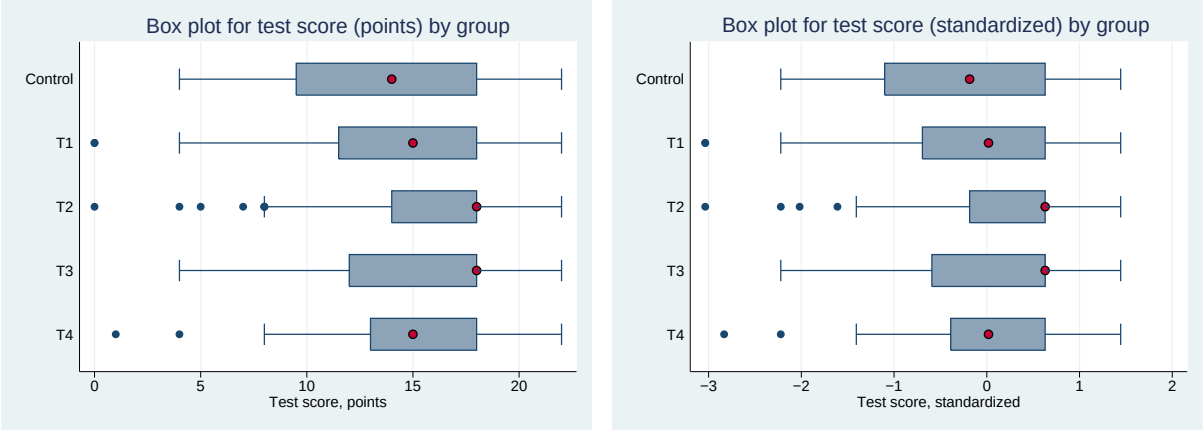
(a) (b) (c)

Figure A.7: Treatment group 4 test score distributions for all students and by gender



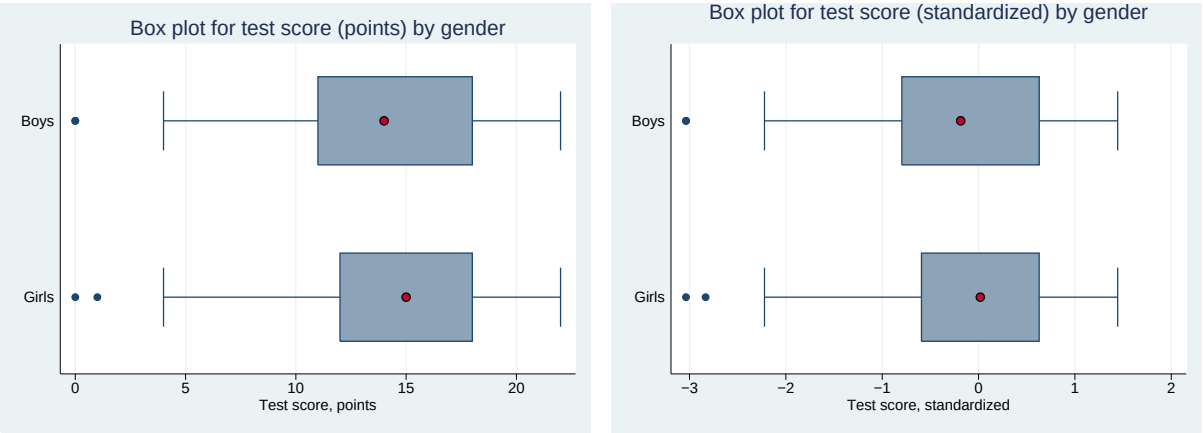
(a) (b)

Figure A.8: Box plot for test score



(a) (b)

Figure A.9: Box plot for test score by group



(a) (b)

Figure A.10: Box plot for test score by gender

Namn: _____ ☐ Flicka

Klass:_____ ☐ Pojke

På detta test kan du få totalt 22 poäng.

<<Information angående testbedömning>>

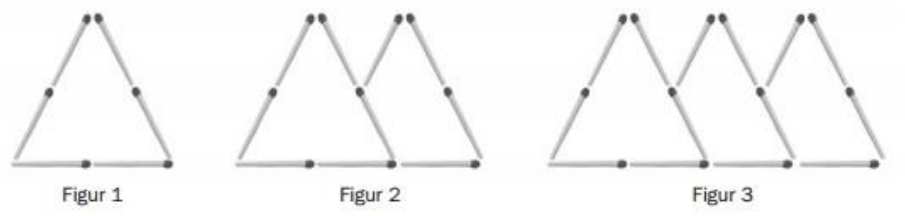
UPPGIFT 1

En kväll började Amanda läsa sin bok på sida 4, hon slutade när hon läst klart sida 11. Nästa kväll läste hon fram till att hon läst klart sida 20, men två av de lästa sidorna innehöll endast bilder.

Fråga: Hur många sidor text läste Amanda totalt på de två kvällarna?

Svar: _____ ___ /4 poäng

UPPGIFT 2



Fråga: Hur många stickor behövs för att bygga en ny figur 4 enligt mönstret ovan?

Svar: _____ ___ /4 poäng

UPPGIFT 3

Emma har två systrar och tre bröder. Emmas bror David har tre systrar och två bröder.

Fråga: Hur många pojkar och hur många flickor finns det i familjen?

Svar: _____ pojkar

_____ flickor ___ /6 poäng

UPPGIFT 4

Räkna ut på det sätt du tycker är bäst.

a) 85,3 – 6,7

Svar:_____

b) 28,5 – 1,3

Svar:_____ ___ /8 poäng

Name: _____ ☐ Girl

Class:_____ ☐ Boy

On this test you can obtain a total of 22 points.

<<Information regarding test assessment>>

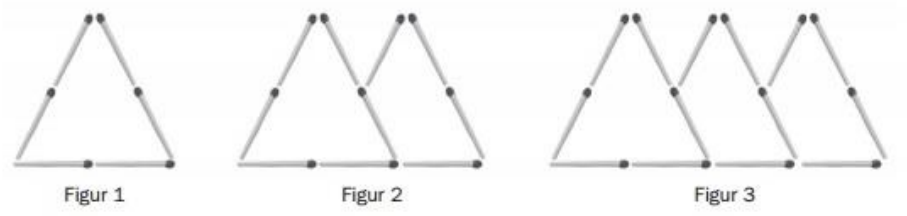
TASK 1

One evening Amanda started to read her book on page 4. She stopped when she had finished reading page 11. The following evening she read until she had finished reading page 20. However, two of the read pages contained only pictures.

Question: How many pages of text did Amanda read in total during the two evenings?

Answer: _____ ___ /4 points

TASK 2



Question: How many matches are needed to build a new figure 4 according to the pattern above?

Answer: _____ ___ /4 points

TASK 3

Emma has two sisters and three brothers. Emma's brother David has three sisters and two brothers.

Question: How many boys and how many girls are there in the family?

Answer: _____ boys

_____ girls ___ /6 points

TASK 4

Calculate the following in the way you prefer.

a) 85,3 – 6,7

Answer:_____

b) 28,5 – 1,3

Answer:_____ ___ /8 points