Seeking Risk or Answering Smart? Heterogeneous Effects of Grading Manipulations in Elementary Schools

Valentin Wagner*

April 3, 2017

Abstract

This paper investigates how grading manipulations affect the quantity and quality of decisions. In a field experiment in elementary schools, 1,377 pupils are randomly assigned to one of three conditions in a multiple-choice test: (i) gain frame (Control), (ii) gain frame with a negative endowment (Negative), and (iii) loss frame (Loss). On average, I find suggestive evidence that pupils in both treatment groups get more points in the test than pupils in the Control Group (coefficients are positive but not significant). Moreover, I find heterogeneous and statistical significant treatment effects when differentiating pupils by ability. High-performers increase performance in both treatment groups, but motivation is significantly crowded out for low-performers only in the Loss Treatment. Additionally, I find that pupils in the Loss Treatment significantly seek more risk—answer more questions—while pupils in the Negative Treatment seem to answer more accurately—increase the share of correct answers. My results have important policy implications, this is, loss framing should not be implemented in (elementary) schools.

Keywords Behavioral decision making, quantity and quality of decisions, framing, loss aversion, field experiment, motivation, education

JEL codes: I20, C93, D03, D80, M54

*Valentin Wagner: Düsseldorf Institute for Competition Economics, wagner@dice.hhu.de. I would like to thank the teachers, parents and pupils who participated in the experiment and the organizers of the Känguru-Wettbewerb for providing the test exercises. I am also grateful for comments and advice from Gerhard Riener, Hans-Theo Normann, Wieland Müller, Axel Ockenfels, Andreas Grunewald, Arnaud Chevalier, Arno Riedl, Sander Onderstal, Heiner Schumacher, Claudia Möllers and participants at the University of Düsseldorf DICE Brown Bag Seminar, the Second Workshop on Education Economics (TIER/LEER in Maastricht), the Third International Meeting on Experimental and Behavioral Social Sciences (Rom), the fifth Workshop "Field Days 2016: Experiments outside the Lab" (Berlin), the seventh International Workshop on Applied Economics of Education (Catanzaro), the 15th TIBER Symposium (Tilburg) and the Annual Conference of the German Economic Association (Augsburg). The usual disclaimer applies.

1 Introduction

Effort is an important prerequisite to achieving externally imposed goals. Managers may set a goal for productivity in the workplace, doctors advise their patient on how much weight to lose, or parents emphasize a GPA target. However, individuals' intrinsic motivation is often too low to achieve these goals. An economist's obvious solution would be the provision of adequate extrinsic financial incentives. While financial incentives can be costly and may have mixed effects on motivation [Gneezy and Rustichini, 2000, Bénabou and Tirole, 2006], there is growing evidence in behavioral economics that non-monetary (recognition) incentives represent an appropriate alternative [Neckermann et al., 2014, Bradler et al., 2016, Kube et al., 2012, Ashraf et al., 2014].[1] Inducing loss aversion to change peoples' behavior tends to be effective and the framing of extrinsic rewards as a loss has been applied to a few field settings [Hong et al., 2015, Armantier and Boly, 2015, Hossain and List, 2012]. These studies demonstrate that the provision of effort is sensitive to incentives framing. Moreover, institutions and governments are increasingly interested in applying insights from behavioral economics into their fields and have installed specialized behavioral insights teams in recent years.[2] It is therefore important to compare the effectiveness of loss framing to other behavioral interventions and to identify for whom loss framing works, along with understanding the underlying mechanisms of effort provision if outcomes depend on multiple inputs, that is, the quality and quantity of decisions.

An ideal setting to test the impact of grading manipulations on the quality and quantity of decisions is within the educational sector by using multiple-choice tests. This testing format creates an environment where decisions have to be taken under uncertainty and performance is dependent on the quality and quantity of answers.[3] It also allows the analysis of heterogeneous framing effects on effort as pupils within a classroom can be differentiated by their initial ability. Moreover, there are not many studies which test the effect of loss framing on performance and motivation in the educational system. Enhancing pupils' motivation is important as it is a key input to excel in the educational system and pupils often invest too little in their own education despite the large returns to education [Hanushek et al., 2015, Card and Krueger, 1992, Card, 1999].[4] To test framing effects is therefore promising as it represents a potential cost-effective and easy-to-implement method to motivate pupils and—to the best of my knowledge—only one paper has applied loss framing on *school-aged* children (secondary schools) so far [Levitt et al., 2016] and no study has yet applied loss framing on children in elementary schools. In particular, testing framing effects on elementary pupils in their last school years in Germany seems to be valuable because the German school system tracks pupils into three different school types—and locks them in tracks throughout middle school—at the early age of 10.[5] Therefore, enhancing pupils' positive attitude towards school (i) might be more effective in younger ages due to complementarities of skill formation at different stages of the education production function [Cunha and Heckman, 2007] and (ii) might influence the tracking decision and thus pupils' future income.[6]

Pupils in elementary schools represent the general population as they are not yet tracked by ability

[1] Wagner and Riener [2015], Springer et al. [2015], Jalava et al. [2015], Levitt et al. [2016] analyze the effectiveness of non-monetary incentives in educational settings.

[2] In 2010, the European Commission set up the "Framework Contract for the Provision of Behavioral Studies (FCPBS)", in 2014 the US government assembled the "Social and Behavioral Sciences Team", the World Bank officially launched its "Global Insights Initiative (GINI)" in 2015 and a number of European countries (United Kingdom, Netherlands, Germany, France and Denmark) installed specialized behavioral insights teams.

[3] Performance in multiple-choice tests can be enhanced by answering more questions (quantity) if the expected number of points when guessing is non negative or by answering questions more accurately (quality).

[4] See Lavecchia et al. [2016] and Koch et al. [2015], for an overview on behavioral economics of education.

[5] A more detailed description of the German tracking system is given in Wagner and Riener [2015].

[6] Results by Dustmann et al. [2016] suggest that pupils in the highest track have 23% higher wages than medium track pupils and completing the medium versus the low track is associated with a 16% wage differential.

and, based on their midterm grades, they can be differentiated into high-, middle- and low-performers.[7] While high-performers are likely to be allocated to the academic track and low-performers to the lower track (preparing for blue collar occupations), middle-ability pupils are at the most risk of being misallocated. Therefore, it is worthwhile to analyze whether different framing manipulations can change the (educational) behavior of all ability groups. Nevertheless, educators might dislike loss framing because pupils could incur psychological or emotional costs.[8] Hence, it is also important to identify alternative ways to increase pupils' motivation. To test loss framing could be appealing for policy-makers as it represents an easy to implement method to potentially boost performance in schools. This is why it is important to inform them about any hidden drawbacks of loss framing, in particular how it works for all pupils of the ability distribution and which domain—risk seeking or accuracy—is mainly affected.

This paper tests whether manipulating the grading scheme improves pupils' performance in a low-stakes 10-item multiple-choice test and compares pupils' answering behavior under three different frames: (1) gain frame, (2) gain frame with negative endowment, and (3) loss frame. Although the test is low-stakes, my results are nonetheless interesting as they give a hint on how grading manipulations could affect the investment in educational inputs (i.e. reading books, doing homework) as they are usually also low-stakes but important to increase academic achievements [Fryer, 2011a, Allington et al., 2010, Kim, 2007]. Moreover, a special focus is on analyzing the effectiveness of grading manipulations on different ability levels (high- and low-performing pupils). To the best of my knowledge, the negative endowment treatment and the differentiation by ability has not been studied previously and represents a major contribution of this paper. Furthermore, the multiple-choice testing format allows me to analyze the impact of framing effects on pupils' risk-seeking behavior and level of accuracy.[9]

The experiment was conducted in 20 elementary schools in Germany among 1377 pupils of grades three and four. The setting of elementary schools allows the analysis of framing effects on heterogeneous ability groups as elementary children are not yet tracked into vocational or academic school types and represent the general population. Pupils were randomized into the *Control Group*, the *Negative Treatment*, and the *Loss Treatment*. In the Control Group and Negative Treatment earning points was framed as a gain. Pupils received +4 points for a correct answer, +2 points for skipping an answer, and 0 points for an incorrect answer.[10] These two treatments differ with respect to pupils' initial endowment—either 0 points or -20 points. Hence, pupils could earn between 0 and 40 points in the Control Group and between -20 and +20 in the Negative Treatment. The intention to endow pupils with a negative amount of points was to make the "passing threshold" more salient. In most exams, pupils need at least half of the points to "pass" the exam or to get a respective grade that signals "pass".[11] In the Loss Treatment earning points was framed as a loss and pupils started with the maximum score (+40 points) but lost -4 points for an incorrect answer, -2 points for a skipped question, and 0 points for a correct answer.

On average, pupils in both treatment groups do not increase the number of point in the test compared

[7] Pupils usually attend the closest elementary school in their neighborhood.

[8] Although some teachers may dislike loss framing, some elementary teachers already use some kind of loss framing in the way they assign "stars and stickers" to pupils. While some teachers give stars for good behavior and reward pupils in case they achieve a predefined amount of stars, other teachers let pupils start with the maximum number of stars but take them away for disruptive behavior. Hence, loss framing is used in education but instead of framing stars as losses, *earning points* is framed as a loss in this study. This information was given informally by some teachers in the run-up to the experiment.

[9] As skipping an answer usually gives a sure (non-negative) number of points, answering a question without certainly knowing the answer is a risky decision. In this study, a risk-neutral individual which does not know the answer is indifferent between answering and skipping a question if the probability of success is 50%.

[10] An incorrect answer is usually punished in multiple-choice tests by deducting points. However, it was important in this experiment that pupils could either only lose or only gain points in order to implement loss and gain framing.

[11] This information was informally given by teachers.

to the Control Group at conventional levels of significance. However, I find heterogeneous framing effects for pupils of different ability levels. While high-ability pupils increase the number of total points in both treatments, low-ability pupils significantly perform worse under the Loss Treatment compared to low-ability pupils in the Negative Treatment and pupils in the Control Group. These results are important, especially for policy-makers who plan to introduce new incentive or grading schemes in schools. Although loss framing might be cost-effective and appears appealing to implement in schools, the experimental results suggest that low-performers—often the main target audience of policy interventions—would be made worse off. Notably, all differences between the treatment groups and the Control Group are driven by a change in (cognitive) effort. The grading scheme of each experimental condition was explained to pupils shortly before they had to take the test. Thus, pupils had no time to study between learning about the grading scheme and the start of the test. This allows a separation of the effort effect from the learning effect. Moreover, pupils in the Loss and Negative Treatment give significantly more *correct answers* compared to pupils in the Control Group. These results seem to be driven by two different mechanisms. In the Loss Treatment, the number of answered questions increases significantly while the share of correctly answered questions does not change. In contrast, the quantity of answers in the Negative Treatment does not significantly differ from the Control Group, while the accuracy of answers significantly increases. This can be interpreted as an increased risk-seeking behavior of pupils in the Loss Treatment and an increase in accuracy of the pupils in the Negative Treatment. Finally, in contrast to Apostolova-Mihaylova et al. [2015], I find no heterogeneous gender effects of loss framing.[12]

The paper is structured as follows. The next section gives an overview about the related literature. The experimental design is described in Section 3 and Section 4 derives theoretical predictions. The data and descriptive statistics are reported in Section 5. Section 6 presents the results which are discussed in Section 7. Section 8 summarizes and concludes.

2 Related Literature

This paper is related to the strand of behavioral literature focusing on loss framing and to the education (economics) literature on grading. Non-monetary incentives to motivate students have received increasing attention by researchers because—in comparison to financial incentives—these kind of rewards are less costly and more importantly, should be widely accepted by teachers, parents and policy makers. Levitt et al. [2016] show that non-monetary incentives (a trophy) work for younger, but not for older children, and that the incentive effect diminishes if the payment of the rewards is delayed. Jalava et al. [2015] find that girls respond to symbolic rewards but that motivation tends to be crowded out for low-skilled students, and Wagner and Riener [2015] test a set of public recognition incentives, showing that self-selected rewards tend to work better than predetermined ones.[13]

Related to grading schemes, Jalava et al. [2015] test the effectiveness of "traditional" criterion-based grading (pupils are graded on an A-F scale according to predetermined thresholds) and rank-based grading system. In the latter, only the top three performers of a class received an A. The authors find that rank-based grading increases the performance of boys and girls, and that rank-based grading also tends to crowd out the intrinsic motivation of low-skilled students.[14] Czibor et al. [2014] investigate the effectiveness of absolute grading and grading on the curve in a high-stake testing environment among university students. The authors

[12]The different findings to Apostolova-Mihaylova et al. [2015] could be due to differences in the subjects' age—university students versus elementary pupils.

[13]See also Bradler et al. [2016], Bradler and Neckermann [2016], Ashraf et al. [2014], Neckermann et al. [2014], Goerg and Kube [2012], Kube et al. [2012] on the effectiveness of recognition and non-financial incentives outside an educational setting.

[14]See also the literature on grading standards mentioned in Jalava et al. [2015].

hypothesize that grading on a curve induces male students to increase their performance when compared to an absolute grading system. They find weak support for this hypothesis and show mainly an increase in performance for the more (intrinsically) motivated male students—female students were unaffected by the grading system. However, there is evidence that rank-based grading could be problematic if ranks are made public. Bursztyn and Jensen [2015] find a decrease in performance if top performers are revealed to the rest of the class, and that signup rates for a preparatory course depend on the peer group composition, that is, to whom the educational investment decision would be revealed. Moreover, educators might dislike rank-based competition between pupils as they are not interested in pupils' relative performance but are more concerned about the individual's learning progress.

Although there is ample evidence on extrinsic rewards and grading schemes, only a few empirical studies have analyzed the effectiveness of framing manipulations among University students and only one study—to my knowledge—on school-aged children. Two very similar studies by Apostolova-Mihaylova et al. [2015] and McEvoy [2016]test whether framing grades of university students as a loss or as a gain affects the course grade at the end of the semester. Students in the treatment group started with the highest possible grade and lost points as the semester progressed while students in the control group started with 0 points and could gain points throughout the semester. After each completed exam or assignment, the students' grades were updated, so that students had the opportunity to follow their increasing or decreasing grades. While McEvoy [2016] finds overall positive effects of loss framing on the final course grade, Apostolova-Mihaylova et al. [2015] find no overall statistically significant effect. However, the authors find heterogeneous gender effects. The final course grade of male students increased while female students got lower grades in the case of loss framing. In contrast, Krawczyk [2011] finds no significant effects of loss framing, subjects are just as risk-averse in losses as they are in gains.[15]

There is little research on framing effects on *school-aged* children. Most similar to my study is the experiment by Levitt et al. [2016] which is the only study testing the loss framing of an extrinsic reward among *school-aged* children in Chicago. The authors provide elementary and high school students with financial ($10 or $20) and non-financial (a trophy) incentives for self-improvement in a low-stakes test. These incentives were announced immediately before the test and were presented as either a loss or gain. In the loss treatment students received the incentive at the beginning of the test and kept it at their desk throughout the test.[16] Levitt et al. [2016] find that immediate paid high financial and non-financial rewards improve performance, and that younger students are more responsive to non-financial rewards. However, they find only suggestive evidence that loss framing improves performance—treatment effects are positive but not statistically significantly different from gain-framed incentives. My study differs in several ways to that of Levitt et al. [2016]: (i) I apply loss framing on *points in a test* and not on an *extrinsic* reward,[17] (ii) loss framing is not only tested against the traditional grading scheme, but *additionally* to a downward shift of the point scale, (iii) loss framing is analyzed for different ability groups, and (iv) the underlying mechanisms of loss framing—impact on quantity and quality of decisions—are examined.

[15] Fryer et al. [2012] analyze whether framing teachers' bonus payments as losses increases the performance of their students. Teachers in the loss frame were paid in advance (lump sum payment at the beginning of the school year) but had to return the bonus if their students did not meet the performance target. The authors find large and statistically significant gains in math test scores for students whose teachers were paid according to the loss frame. The size of gains was equivalent to increasing teacher quality by more than one standard deviation.

[16] Students had to sign a sheet confirming receipt of the reward and were asked to return it in case of missing improvement.

[17] Framing points as gain or loss should help to maintain a "natural" testing environment, as pupils usually do not get extrinsic rewards for performance in a test.

3 Experimental Design

The experiment was conducted in 20 elementary schools with a total of 71 school classes in the federal state of North Rhine-Westphalia (NRW), Germany. During May and November 2015, 1377 pupils in grades three and four participated.[18] With the semester report in grade four, parents receive a transition recommendation to which school type—academic or vocational track—to send their child. This recommendation is given by the elementary school teacher and is based on i) talent and performance, ii) social skills and social behavior, and iii) motivation and learning virtues [Anders et al., 2010]. However, parents in NRW have the choice to which type of secondary school they want to send their children, regardless of the school recommendation. Nevertheless, depending on their capacity, secondary schools can decline applications.[19] Hence, policy interventions to boost pupils' performance in grades three and four might have long-lasting effects, as these grades are important stages for the recommendation decision and promotion within the German school system.

3.1 Selection of Schools and Choice of Testing Format

Selection of Schools In total, 221 elementary schools in the cities of Bonn, Cologne and Düsseldorf, which represent about 7.7% of all elementary schools in NRW were contacted based on a list that is publicly available from the Ministry of Education of NRW. The first contact was established via Email on April 7, 2015 and a second mailing followed on August 3, 2015 (at the end of the summer holidays). About 19% of all contacted schools responded, and 50% (21 schools) of these schools replied positively and agreed to a preparatory talk.[20] In these talks, the experimental design was explained to at least one teacher and lasted about 20-30 minutes. Finally, 20 schools totaling 71 classes participated in the experiment. One school initially agreed to participate and received all experimental instructions and testing material but finally did not carry out the experiment. The reasons are not known as the school did not respond to any mailing afterwards. Additionally, one teacher of another school did not manage to write the test on time due to illness.

Multiple-Choice Test The mathematical test in this experiment consisted of 10 multiple-choice pen-and-paper questions and represented a compilation of old age appropriate questions of the *"Känguru-Wettbewerb"*.[21] The "*Känguru-Wettbewerb*" is administered once a year throughout Germany and uses age appropriate test questions. Pupils had 30 minutes to answer all the questions so that the test could be taken in a regularly scheduled teaching hour.[22] The problems and the answer options were presented on three question sheets and points could be earned according to the treatment specifications (see Table 1). There were five answering possibilities with only one correct answer per question, and pupils had to mark their answers on the same sheet. To minimize cheating [see Armantier and Boly, 2013, Behrman et al., 2015, Jensen et al., 2002], the order of questions was changed within the class. To fulfill privacy and data protection requirements, each test and questionnaire received a test identification number, so that pupils did not have

[18] Elementary school in Germany runs from grade one at the age of 6 to grade four at the age of 9 or 10.

[19] Criteria for the admission decisions that may be used by the school principal are the number of siblings already attending the school, balanced ratios of girls and boys, distance to school and/or a lottery procedure (see http://www.schulministerium.nrw.de/docs/Recht/Schulrecht/APOen/HS-RS-GE-GY-SekI/APO_SI-Stand_-1_07_2013.pdf).

[20] Non-participating schools which replied to the request declined participation due to a number of other requests of researchers or limited time capacities.

[21] The *Känguru-Wettbewerb* consists of 24 items and working time is 75 minutes. Hence, 10 questions were chosen in the experiment to adjust for the shorter testing time of 30 minutes.

[22] A regular teaching hour in Germany lasts for 45 minutes.

to write down their names. This procedure is similar to the one of evaluations of learning processes which are regularly carried out in various subjects. Furthermore, parents had to sign a consent form ("opt-in").[23]

3.2 Treatments

The following three treatments were designed to analyze the effectiveness of different grading schemes on pupils' performance: the Control Group (Control), the Loss Treatment (Loss), and the Negative Treatment (Negative). The test was announced one week in advance in all treatments, and the preparatory material for pupils was distributed in the same lesson. During the preparation week, teachers were not allowed to actively prepare pupils for the test.[24] The grading scheme differed across treatments and was announced to pupils on the testing day shortly before the test started. This design therefore allows the measuring of a pure effort effect and no learning because pupils had no time to study after the grading scheme was communicated.[25] Any treatment effects can therefore be attributed to pupils exerting more effort during the test and not to a learning effect—for example, pupils spending more time on test preparation.

Control Group Pupils in the Control Group started the test with 0 points which is the "traditional" way in Germany. For each correct answer, pupils earned +4 points, 0 points for a wrong answer, and +2 points when they skipped a question. Hence, pupils could never lose a point in the Control Group, and consequently could earn between 0 and +40 points. Note that a sure gain of +2 points for skipped answers increases the cost of guessing under uncertainty. Risk-neutral individuals who maximize the expected number of points but do not know the correct answer and cannot exclude a wrong answering choice, are indifferent between answering and skipping the question if the probability of finding the right answer is 50%.

Loss Treatment To implement loss aversion, pupils were endowed with the maximum score of +40 points upfront, but subsequently could only lose points. Pupils earned -4 points for a wrong answer, -2 points for skipping a question, and 0 points for a correct answer. Like pupils in the Control Group, they could earn between 0 and +40 points.

Negative Treatment In the Negative Treatment, earning points was framed in the same manner as in the Control Group. Pupils earned +4 points for a correct answer, 0 points for a wrong answer and +2 points for skipping a question. The only difference between the Negative Treatment and Control Group was that pupils started the test with -20 points.[26] Thus, pupils could earn between -20 and +20 points. Usually pupils in Germany are graded on a strict scale and have to score at least half of the points to "pass" the exam. Hence, this treatment intended to make the threshold of passing more salient.

In many multiple-choice testing formats, pupils can gain points for correct answers and lose points for incorrect ones. However, to be able to test loss framing, it was necessary that pupils could only gain points in the Control Group, and only lose points in the Loss Treatment. Notice that pupils in the Control Group and Loss Treatment who give the same number of correct answers and skip the same number of questions earn the same amount of total points in the test. This is also true for pupils in the Negative Treatment if the negative endowment of -20 points is taken into account.

[23] The experimental design excludes the possibility of non-random attrition as the same consent form was given to the treatment and control groups. Hence, selection into treatments is not a major issue. Attrition is discussed in detail in Section 5.1.

[24] Teachers answered questions concerning the preparatory exercises only if pupils asked on their own initiative.

[25] See also the experimental design by Levitt et al. [2016] for isolating the effort effect from the learning effect.

[26] Pupils in grades three and four already learned addition and subtraction with numbers from 0 up to 100.

One concern of implementing loss framing or a negative endowment among third and fourth grader participants could be how well they understood the treatments. However, pupils in these grades have already learned addition and subtraction with numbers from 0 up to 100 and it is reasonable to assume that they have an intuitive grasp that negative points are worse than positive points. Nevertheless, if pupils do not understand the treatments then pupils' behavior in the treatment groups should not differ from the behavior of pupils in the Control Group. However, as can be seen in section 6, pupils respond to loss framing and behave as predicted by prospect theory by seeking-more risk. Table 1 gives an overview of the treatment conditions.

Table 1: Treatment Overview

	Starting Points	*Correct Answer*	*Skipped Answer*	*Wrong Answer*	*Minimum Points*	*Maximum Points*
Treatments						
Control	0	+4	+2	0	0	+40
Loss	+40	0	−2	−4	0	+40
Negative	−20	+4	+2	0	−20	+20

Note: This table displays the number of points pupils received for a correct, wrong or skipped answer as well as the amount of starting points and the minimum and maximum number of total points separately for each treatment.

Randomization

Randomization was performed using a block-randomized design.[27] Blocked on grade level within schools, classes were randomized into the Control Group, Loss Treatment, or Negative Treatment. Hence, all pupils within the same class were randomized into the same treatment. The randomization procedure ensured that the Control Group and either the Loss or the Negative Treatment were implemented within each grade level of a school participating in the experiment with two classes.[28] The Loss and Negative Treatment were implemented simultaneously for schools participating with three or more classes of the same grade level.

Table 7 in Appendix A.1 shows the randomization of treatments and reports on the number of participants, average number of correct answers, and average points by treatment group i) for the full sample, and ii) separately for boys and girls. Table 8 in Appendix A.1 presents randomization checks adjusting for multiple hypothesis testing [see List et al., 2016]. On average, the variables do not differ from the Control Group at conventional levels of statistical significance. This indicates that the randomization procedure was successful. However, teachers seem to be less experienced on average in the Negative Treatment. Having less experienced teachers could have a negative effect on pupils' performance and therefore would underestimate positive treatment effects. Furthermore, there is a significant difference between the Control Group and both treatment groups with respect to the timing of writing the test. However, I control for these differences in the statistical analysis. Differences in baseline scores and concerns about non-random self-selection are discussed in subsection 5.1.

Participants are on average 9.10 years old and have 0.79 older siblings. Of the total, 48.80% are female and 78.44% speak German at home. The average midterm grade in mathematics is 6.48 on a scale from 1 to 15, where 1 is the highest and 15 is the lowest grade.[29]

[27] See Duflo et al. [2007], Bruhn and McKenzie [2009] regarding the rationale for the use of randomization.

[28] There were only two schools in which one class participated.

[29] Midterm grades in Germany usually take on values 1+, 1, 1−, 2+, 2, 2−, . . . 6−. However, to better deal with these grades in the analysis, I code midterm grades from 1 to 15. Midterm grade 15 (= 5-) is the lowest grade as no child had a grade below this.

3.3 Implementation[30]

Researchers were never present in the classroom to maintain a natural exam situation within the classroom. Therefore, teachers got detailed instructions in the run-up of the experiment. Each school was visited once during the preliminary stage of the experiment. In this meeting, the exact schedule and procedure of the experiment was described and teachers' questions were answered. Each teacher received the instructions again in written form close to the start of the experiment. In total, two envelopes were subsequently sent to the teacher. The first envelope was distributed at the beginning of the experiment—the moment a school agreed to participate—and contained instructions regarding the announcement of the test, preparatory material for pupils and consent forms for parents (see Appendix). At this point teachers, got to know their treatment group but were not yet allowed to communicate it to pupils. It was necessary to tell teachers their treatment group in advance to give them the opportunity to ask questions of clarification. Two to three days before the test date, teachers received the second envelope containing the tests, detailed instructions for implementations on the test day and a list in which teachers were asked to enter pupils' midterm grades and the corresponding test-id numbers.[31] It was important to send the tests in a timely manner in order to reduce the risk of intentional or unintentional preparation of pupils by teachers. Teachers and pupils answered a questionnaire at the close of the experiment.

It was common to all treatments that teachers were asked to choose a suitable testing week in which no other class test was scheduled for which pupils had to study. Teachers announced the test one week in advance and distributed the preparatory questions with attached solutions as well as the consent forms to be signed by parents. The teachers clarified that pupils' performance will be evaluated and that pupils will get a grade but that this grade does not count for the school report. They did so in the framework of an evaluation of pupils' achievements which demonstrates their skills during a school year. Pupils had 30 minutes to answer all the test questions and filled out a questionnaire that was attached to the end of the test. The tests were corrected centrally by the researcher, graded by teachers and pupils received their result shortly after (section 6.1 presents arguments why it is interesting to examine treatment effects on low-stakes tests.)

At the testing day, teachers explained in detail how pupils could earn points shortly before the test started and the introductory text at the top of the tests varied by treatment:

Control:

"1. Please do not write your name on the test. For each task, there are 4 wrong and 1 correct answers. Please write your answers in the boxes.

2. The highest possible score is 40, the lowest 0.

3. You start with 0 points. If a correct answer is written, you get +4 points. You get +2 points if no answer is given and 0 points if an incorrect answer is written."

Loss:

"1. Please do not write your name on the test. For each task, there are 4 wrong and 1 correct answers. Please write your answers in the boxes.

[30] The implementation of the experiment is similar to Wagner and Riener [2015].

[31] Due to data privacy reasons, each pupil got a test-id number so that researchers could not infer pupils' identity.

2. The highest possible score is 40, the lowest 0.

3. You start with the maximum number of points. This means you have 40 points at this point. However, you lose 4 points if an incorrect answers is written and you lose 2 points if no answers is given. If a correct answer is written, you lose no points."

Negative:

1. Please do not write your name on the test. For each task, there are 4 wrong and 1 correct answers. Please write your answers in the boxes.

2. The highest possible score is +20, the lowest -20.

3. You start with the minimum number of points. This means you have -20 points at this point. However, if a correct answer is written, you get +4 points. You get +2 points if no answer is given and 0 points if an incorrect answer is written."

4 Theoretical Predictions

The objective of this paper is to test whether loss framing of points in a multiple-choice test increases performance of elementary children. I therefore consider a model based on prospect theory in which pupils derive gain-loss utility relative to a reference point [Kahneman and Tversky, 1979]. Pupils' decision problem while answering multiple-choice questions can be thought of in two ways: First, pupils have to decide whether they want to answer or to omit the question which I will refer to as the risk-taking decision. Second, if pupils have decided to answer the question, they have to exert (cognitive) effort which I will refer to as the effort decision. Grading manipulations can therefore increase (or decrease) the number of points in the test due to a change in the risk-taking decision, a change in the effort decision or a mixture of both. In the following, I present a simple model of pupils' decision problem which consists of two stages. Stage 1 models the risk-taking decision and is similar to the model presented in Krawczyk [2011]. Stage 2 models the effort decision and boroughs heavily from the model in Levitt et al. [2012][32]. Furthermore, I assume that the risk-taking and effort decisions are not influenced by the answering history, as pupils do not get direct feedback about the correctness of their answers.

Stage 1 (risk-taking decision) Pupils read the text of the multiple-choice question and thereby form (ad hoc) a minimum individually perceived probability of selecting the correct answer (μ).[33] As omitting a question gives 2 points with certainty, pupils answer the question only if μ is sufficiently high. Furthermore, pupils derive utility v from points p which is the value function of prospect theory and is convex in losses and concave in gains. $p_{correct}$, p_{wrong} and $p_{omitted}$ are the points pupils get for a correct, wrong and omitted answer and differ between treatments as outlined in table 1. Furthermore, $v(0)$ is normalized to 0 and w is a probability weighting function as in Kahneman and Tversky [1979] (for a better readability the indices *gain* and *loss* are attached to the respective weighting function, indicating whether pupils are in a gain framing or in a loss framing condition).

[32] The theoretical model is presented in the working paper version but not in the finally published paper which is why I cite the working paper here instead of the published article.

[33] μ can also be thought of as a "feeling" of rather finding the correct answer or of rather answer the question wrong.

Hence, pupils in the gain frame condition (Control Group) are indifferent between answering and omitting the question if

$$w_{gain}(\mu)v(p_{correct}) + w_{gain}(1-\mu)v(p_{wrong}) \;=\; v(p_{omitted})$$

taking the points of the gain framing condition it follows $w_{gain}(\mu)v(4) = v(2)$. If—for simplicity—linear probability weighting is assumed, pupils are indifferent between answering and omitting the question whenever $\mu = \frac{v(2)}{v(4)}$, and if v is concave for gains, this implies that $\mu > \frac{1}{2}$. Pupils in the loss framing condition (Loss Treatment) are indifferent between answering and omitting the question if

$$w_{loss}(\mu)v(p_{correct}) + w_{loss}(1-\mu)v(p_{wrong}) \;=\; v(p_{omitted})$$

which is

$$w_{loss}(1-\mu)v(-4) \;=\; v(-2)$$

With linear probability weighting, pupils in the loss framing condition answer the question if $1-\mu = \frac{v(-2)}{v(-4)}$. If v is convex for losses, than $\frac{v(-2)}{v(-4)} < \frac{1}{2}$ and hence the minimum individually perceived probability of selecting the correct answer is lower in the Loss Treatment than in the Control Group. Consequently, less questions will be omitted if pupils code their situation in the domain of losses. The risk-taking decision for pupils in the Negative Treatment is less clear and should depend on whether they adjust their reference point to the incurred loss or not. If their reference point is determined by the status quo [Thaler and Johnson, 1990], they would code their situation in the domain of gains and their answering behavior should not differ from pupils in the Control Group. However, if pupils in the Negative Treatment do not adjust their reference point to the new endowment, they are in a loss domain and should answer similar to pupils in the Loss Treatment. Notice, that the change in pupils' risk-taking decision between the gain and loss framing condition is orthogonal to pupils' risk preferences. Risk-loving pupils have a lover minimum perceived probability of finding the correct answer than risk-averse pupils, i.e. $\mu^{risk-loving} < \mu^{risk-neutral} < \mu^{risk-averse}$, but loss framing would always lead to an increase in risk taking compared to the gain framing condition.[34]

Hypothesis 1 *Pupils in the loss frame condition answers more questions than pupils in the gain frame condition due to loss aversion. Pupils in the Negative Treatment answer more questions if they do not adjust their reference point to the new (negative) endowment.*

Stage 2 (effort decision) If pupils have decided to answer the question in stage 1, they have to decide on the (cognitive) effort which directly effects the probability of finding the correct answer. I borough heavily from the model described by Levitt et al. [2012] in which pupils make reference-dependent effort decisions and derive benefits and costs from answering as follows:

$$v(e,\, p,\, p^r) = \pi(e)[u(p_{correct}) + R(p_{correct},\, p^r)] + [1-\pi(e)][u(p_{wrong}) + R(p_{wrong},\, p^r)] - c(e)$$

where pupils answer the question correctly with probability π and get $p_{correct}$ points (and answer the question incorrectly with $1-\pi$ and get p_{wrong}points), u is the utility over points, R is the value function of

[34] Risk-averse pupils in the loss frame seek more risk than risk-averse pupils in the gain frame and risk-loving pupils in the loss frame will seek more risk than risk-loving pupils int the gain frame.

prospect theory and c is the cost of effort e. Let $\pi(.)$ be increasing and concave in e, $u(.)$ be increasing and concave in p, $c(.)$ be increasing and convex in e, and $u(0) = 0$. Furthermore, utility is derived in relation to a reference point $p^r \epsilon (0,\ 4)$ and $R(.)$ and defined as:

$$R(p,\ p^r) = \begin{cases} g(p - p^r), & if\ p \geq p^r \\ h(p - p^r), & if\ p < p^r \end{cases}$$

where g is increasing and concave, h is increasing and convex and $g(0)$ is normalized to 0. The objective function for pupils in the *gain frame condition* is then

$$\frac{max}{e} \pi(e)[u(4) + R(4,0)] + [1 - \pi(e)][0 + R(0,0)] - c(e)$$

and the objective function for pupils in the *loss frame condition* is

$$\frac{max}{e} \pi(e)[u(4) + R(4,4)] + [1 - \pi(e)][0 + R(0,4)] - c(e)$$

the respective first order conditions are $c^{'} = \pi^{'}(e)[u(4) + g(4)]$ and $c^{'} = \pi^{'}(e)[u(4) - h(-4)]$. If losses are felt more strongly then gains $-h(-4) > g(4)$, then optimal effort will increase if points are framed as loss rather than gains. Again, the effort decision of pupils in the Negative Treatment depends on whether they adjust to the new endowment or not. If they adjust their reference point to the new endowment, the effort decision does not differ from pupils in the Control Group and if they do not adjust their reference point their effort decision is similar to those pupils in the Loss Treatment.

Hypothesis 2 *Pupils exert more (cognitive) effort in the loss frame condition compared to pupils in the gain frame condition and hence increase their probability to answer correctly—increase the share of correct answers.*

5 Data and Descriptive Statistics

Data on pupil and teacher level are questionnaire based and compared to data in NRW. The most important control variable is pupils' last midterm grade in math to be able to control for pupils' baseline performance. Midterm grades have the advantage that they are reported by teachers and can be treated as exogenous in the analysis because they were given to pupils before teachers learned about the experiment. Midterm grades in Germany combine the written and verbal performance of pupils wherein the written part has a larger influence on the final course grade and should be correlated with pupils' true ability; thus, these grades are a good—also not perfect—measure of mathematical ability. Further control variables at the pupil-level I will use to derive my results in Section 6 are gender, parents' education and a dummy whether pupils are in grade three or four. The latter variable controls for pupils' age and educational level simultaneously. Parents' educational level is captured by the number of books at home (see Wößmann [2005], Fuchs and Wößmann [2007] for an application in PISA studies).

Control variables at the classroom-level are teachers' working experience, the number of days between the test and the next holidays, and an indicator whether the test was written before or after the summer holidays. It seems that there is a common understanding in the literature that unobserved teacher characteristics may be more important than observed characteristics. Among the observable teacher characteristics, many studies find a positive effect of teachers' experience on pupils' achievement [Harris and Sass, 2011, Mueller, 2013].

The number of days until the next holidays is included as pupils' academic motivation could change as the semester progresses [Corpus et al., 2009, Pajares and Graham, 1999]. Pupils who write the test close to the start of the holidays could be less motivated to exert effort than pupils who write the test at the beginning of the semester.[35] It was also necessary to include a dummy controlling whether the test was written before or after the summer break as the summer break marks the beginning of the new school year. Controlling only for the school grade would neglect the fact that pupils in grade four before the summer break are one year ahead in the teaching material than pupils in grade four after the summer break.

Table 2 compares the descriptive statistics to the actual data in NRW. Although representativeness of the sample for the school population in NRW cannot be claimed, the data are consistent with key school indicators.[36] 1.333 observations were included in the final analysis; 44 observations were dropped because of missing values.[37]

Table 2: Comparison of characteristics: Experiment vs. North Rhine-Westphalia (in %)

	Experimental Data	*North Rhine-Westphalia*
Proportion Female	48.80	49.19
Proportion Pupil German	62.89	56.40
Class Size	24.85	23.20
Proportion Teacher Female	94.29	91.27

Note: This table compares characteristics of the pupils in the experiment with the same indicators in NRW. Cell entries represent percentages of key school indicators. NRW school data are taken from the official statistical report of the ministry of education for the school year 2014/2015 (see https://www.schulministerium.nrw.de/docs/bp/Ministerium/Service/Schulstatistik/Amtliche-Schuldaten/StatTelegramm2014.pdf). *Proportion Female* is the share of females, *Proportion Pupil German* is the share of pupils without migration background, *Class Size* is the average number of children in a class and *Proportion Teacher Female* is the share of female teachers.

5.1 Self-Selection

Parents had to give their consent that their child was allowed to participate in the experiment and that teachers were allowed to pass on pupils' test as well as midterm grades to the researcher.[38] Hence, before comparing the performance of pupils in the two treatment groups to the Control Group, concerns related to non-random attrition need to be alleviated. If attrition is associated with the outcomes of interest, then the results could lead to biased conclusions. Nevertheless, biased outcomes are unlikely if response probabilities are uncorrelated with treatment status [Angrist, 1997].

There are several reasons for attrition: (i) pupils are sick on the testing day, (ii) pupils have lost or

[35] In total there were two holidays during the experiment (summer and autumn).

[36] The difference in "Proportion Pupil German" could be due to the fact that the experiment was conducted only in schools of larger cities.

[37] Missing values were mainly the result of incomplete pupil questionnaires. There are 3 missing values for the last midterm grade and 41 for pupils' gender.

[38] This is a necessary legal prerequisite in NRW to conduct scientific studies with under-aged children (see https://www.schulministerium.nrw.de/docs/Recht/Schulrecht/Schulgesetz/Schulgesetz.pdf and http://www.berufsorientierung-nrw.de/cms/upload/BASS_10-45_Nr.2.pdf).

forgotten the signed consent form, (iii) parents forgot to sign the consent form on time but actually agreed, or (iv) parents intentionally did not give their consent. I cannot disentangle the reasons for attrition because the data set contains information only about those pupils who participated in the test and handed in the consent form in time. Most importantly, the experimental design excludes the possibility of strategic attrition as all parents got the same consent forms in the treatment and control groups, and hence received the same information about the experiment. Therefore, parents did not get to know which treatment was implemented in the classroom of their child.

There is also no support for non-random attrition in the data. Table 9 in Appendix A.2 reports on the average number of absent pupils and the average ability (midterm grades) of the class by treatment. Comparing treatment groups to the Control Group shows that fewer pupils are absent on average in the Loss Treatment (4.27 vs. 4.13; t-test yields a p-value of 0.909) but that a higher share of pupils is absent in the Negative Treatment (4.27 vs. 6.27; p = 0.175). The average ability level seems to be lower in the Loss Treatment (6.49 vs. 6.68; p = 0.572) and higher in the Negative Treatment (6.49 vs. 6.26; p = 0.478) as compared to the Control Group. However, these differences in midterm grades are small in size. Midterm grades in the dataset are coded on a scale from 1 to 15, where 1 is the highest and 15 the lowest grade (e.g. a midterm grade of 6 represents a B+ and a midterm grade of 7 equals a C-). Nevertheless, these small differences in midterm grades are controlled for in the regression analysis. Moreover, none of the observed differences (average class ability and rate of absenteeism) are statistically significant. Results should therefore not be biased by non-random selection.

6 Experimental Results

The result section is organized in the following way. First, I discuss why it is valuable to analyze grading manipulations in a low stakes test. Second, the effectiveness of framing on the number of points is analyzed using count data regression models (ordinary least square regressions are presented in Table 15 in Appendix A.4). Thereafter, treatment effect estimates are presented for the number of correct answers and omitted questions using poisson and negative binomial regression models. Ordinary least square regression is then used to estimate treatment effects for the share of correctly given answers—the number of all correct answers divided by the number of given answers (correct + incorrect). Finally, I differentiate pupils by ability and gender. The results are discussed thereafter.

6.1 Low stakes testing

It was not possible to implement the experiment in a high stakes testing environment—test score counts for pupils' overall grade—due to the institutional setting and teachers' resistance.[39] Hence, the multiple-choice test is a low stakes test which is also the case for PISA and other standardized comparative tests (i.e. VERA, IGLU, TIMSS). Although it would be interesting to analyze the effectiveness of grading manipulations in both, low and high stakes test, I opted for the former for two reasons: First, although stakes are low for pupils, they might by high in standardized test for schools and teachers. Second, the stakes of incentivizing educational inputs (i.e. reading books) and the stakes of incentivizing performance in a test which does not count towards the final course should be of similar magnitude. Hence, results on the effectiveness of grading manipulations in a low stakes can be informative to educational policy makers on how to increase educational inputs. This is

[39]Teachers did not agree that the test performance counts for the final grade—because contrary to regular exams—the multiple-choice test of the experiment does not test recently learned curricular content.

important as incentivizing educational inputs could be more effective than incentivizing educational outputs [Sadoff, 2014, Fryer, 2011b, Allington et al., 2010, Kim, 2007].[40]

6.2 Framing and test performance

The main outcome variable of interest is the number of correct answers in the test and represents count data. As discussed later, it is also interesting to analyze the effects of grading manipulations on the number of correct answers, omitted answers and the share of correct answers. The identification of the average treatment effects—differences between treatment and Control Group means—relies on the block randomization strategy. To estimate the causal impact of framing on pupils' performance, treatment effects are estimated by applying count data models. Control variables on pupil- and class-level are included as well as school fixed effects.[41] Standard errors are clustered on class-level—which is the level of randomization. Therefore, I estimate the following Poisson model:

$$E(Points_i) = m\left(\beta_0 + \beta_1 Treatment_i + \beta_2 Midterm_i + \gamma P_i + \mu C_i + \delta School_i\right) \tag{1}$$

$m(.)$ is the mean function of the Poisson model. $Points_i$ is the number of points by pupil i, $Treatment_i$ indicates the respective treatment, $Midterm_i$ is the grade in math on the last semester report, P_i is the vector of pupil-level characteristics, C_i a vector of class-level covariates (covariates are described in detail in Section 5) and $School_i$ controls for school fixed effects. A linear model (OLS) is estimated as a robustness check; the results do not change neither in significance nor size (see Table 15 in Appendix A.4).

Table 3 presents estimates of the average treatment effects for the Loss Treatment and Negative Treatment. The dependent variable is the number of points in the test (in marginal units) with standard errors clustered on class-level. The first column presents estimates without controls but school fixed effects. The second column controls for class characteristics and the third column controls for pupil characteristics. The fourth column controls for both class and pupil control variables and is the specification of interest.

When controlling only for school fixed effects or for school fixed effects and additionally classroom covariates (columns 1 and 2), pupils which are exposed to a negative endowment of points significantly increase the number of final points compared to pupils in the Control Group. In contrast, coefficients are negative, small and insignificant for pupils in the Loss Treatment. However, the significant result in the Negative Treatment vanishes when controlling for pupil-covariates, in particular, controlling for pupils' baseline performance (midterm grades). Hence, I find *suggestive* evidence that both treatments increase overall performance as coefficients on the number of total points are positive (as expected); however, this result is not definitive.

Result 1 *Loss framing and a negative endowment tend to outperform a "traditional" grading.*

[40] There are several reasons why pupils should be motivated to put some effort into a low stakes test. First, grades (and ranks) themselves have an incentive effect [see Koch et al., 2015, Lavecchia et al., 2016, and the literature mentioned therein]. Second, pupils might want to signal good performance to parents or the teacher [see Wagner and Riener, 2015] and third, giving feedback on performance allows for social comparison within the classroom (Bursztyn and Jensen [2015] show that pupils' investment decision into education differs based on which peers they are sitting with and thus to whom their decision would be revealed).

[41] Furthermore, there has not been a change of the teacher between the midterm grade and the test.

Table 3: Treatment Effects - Total Points in Test

	(1)	(2)	(3)	(4)
Treatments				
Loss	−0.007	−0.037	0.358	0.178
	(0.739)	(0.716)	(0.631)	(0.595)
Negative	1.604*	1.545**	0.826	0.846
	(0.875)	(0.785)	(0.807)	(0.654)
Controls				
ClassCov	No	Yes	No	Yes
PupilCov	No	No	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes
N	1333	1333	1333	1333

Note: This table reports the marginal effects of a negative binomial regression including school fixed effects. Dependent variable: total number of points in test. Covariates: last midterm grade, gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. 44 observations are dropped due to missing values. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

While the effectiveness of grading manipulations on the number of points is of interest for education policy makers, the number of correct answers might be of greater interest from the teachers' perspective. The number of total points is uninformative for teachers as points can be gained either by answering correctly or by skipping questions. For example, 20 points can be achieved by either giving 5 correct and 5 incorrect answers or by skipping 10 questions. However, teachers want to learn about whether pupils are able to answer the question correctly to better tailor their teaching to pupils' needs. Therefore, teachers might be rather interested to know how the treatments effect the number of correct answers than the number of total points.

Pupils in the Loss Treatment as well as pupils in the Negative Treatment increase the number of correct answers compared to pupils in the Control Group. These findings are statistically significant at conventional levels. Pupils in the Loss Treatment give on average 0.436 ($p = 0.002$) more correct answers which is an increase by about 11.2% compared to the performance of pupils in the Control Group. Similarly, pupils in the Negative Treatment increase their performance by about 8% (marginal effect: 0.309; $p = 0.029$). The difference between the Loss and Negative Treatment is statistically not significant.[42]

[42]The change in significance levels between column (1) and (3) is driven by controlling for pupils' past performance.

Result 2 *Loss framing and a negative endowment increase significantly the number of correctly solved questions.*

Table 4: Treatment Effects - Number of Correct Answers

	(1)	(2)	(3)	(4)
Treatments				
Loss	0.332	0.376*	0.456***	0.436***
	(0.217)	(0.198)	(0.157)	(0.140)
Negative	0.500**	0.516**	0.265	0.309**
	(0.237)	(0.213)	(0.193)	(0.143)
Controls				
ClassCov	No	Yes	No	Yes
PupilCov	No	No	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes
N	1333	1333	1333	1333

Note: This table reports the marginal effects of a Poisson regression including school fixed effects. Dependent variable: number of correct answers. Covariates: last midterm grade, gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. Standard errors are reported in parentheses and clustered on classroom-level. 44 observations are dropped due to missing values. The number of clusters is 71. Robustness checks with OLS regressions show similar results (see Table 15 in the Appendix).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Seeking Risk or Answering Smart? It is crucial for educators to explore the underlying channels—risk-seeking or cognitive effort—through which loss framing increases performance before implementing it in a large scaled intervention. Treatment effects on the number of correct answers are significantly positive in the Loss and Negative Treatment. One reading of these results could be that pupils exert more cognitive effort or—as prospect theory would predict—pupils increase their willingness to choose risky lotteries. Thus, the results could be driven by an increase in the willingness to answer risky multiple-choice questions rather than exerting more cognitive effort.[43]

The multiple-choice testing format allows to identify which mechanisms (effort or risk-seeking) increases the number of correct answers in the Loss and Negative Treatment. For each test item, pupils have to decide whether they want to answer or skip the question. Answering a question without certainly knowing the correct answer is a risky decision and gives—in expected value—a positive number of points only if the probability to answer the question correctly is above 50%. Therefore, differences in the number of skipped questions

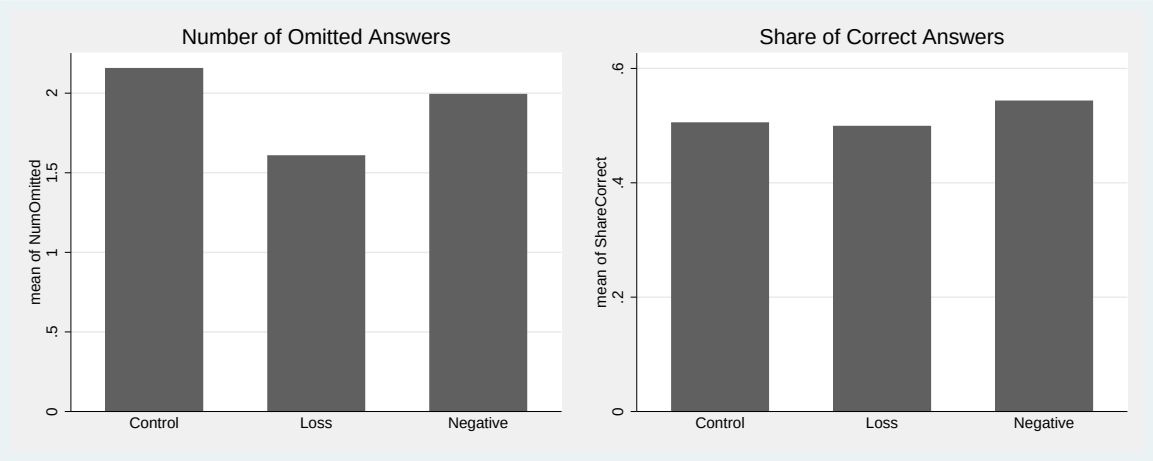
[43]Risky multiple-choice question refers to a test question where the answer is unknown and thus answering this question is a decision under uncertainty.

between the Control Group and the treatments groups would be an indication of a change in risk-seeking behavior. As the test is low stakes, the number of skipped questions can also be viewed as a proxy of effort, i.e. whether pupils even bother to fill in the questions [Zamarro et al., 2016]. However, this can be checked by the different ordering of test questions between tests and is discussed in more detail in section 7. I do not find evidence that the number of omitted questions measures effort. Prospect theory predicts that pupils become more risk-seeking if gambles are framed as a loss [Kahneman and Tversky, 1979] and hence, pupils are likely to become more risk-seeking in the Loss Treatment which means that they skip fewer questions. Whether the risk-seeking behavior changes in the Negative Treatment is less clear as earning points is framed as a gain. Nevertheless, pupils may become more risk-seeking in order to avoid a negative number of total points in the test or because they have more pessimistic beliefs about the grade they would get with a negative score. Another variable of interest is the share of correct answers because it can be interpreted as a measure of "accuracy". The term accuracy refers to the case in which pupils exert more cognitive effort—increasing the probability of answering correctly. In order to increase the number of correct answers, pupils could either take the risky-lottery and answer more questions or they could answer the same number of questions but increase the probability of success by exerting more cognitive effort. Thus, if pupils answer more questions but do not increase the share of correctly given answers, this would be an indication that they became more risk-seeking. On the other hand, if they answer the same amount of questions but increase the share of correct answers would be an indication that they increase their accuracy level. It is also conceivable that both treatment groups increase the risk-seeking behavior and the accuracy level simultaneously.

The analysis of descriptive data—Figure 1—suggests that pupils in the Control Group skip more questions than pupils in the Loss Treatment (2.155 vs. 1.607, $p < 0.001$) while the share of correct answers does not differ between these two groups (0.5049 vs. 0.4988, $p = 0.709$). In contrast, the difference in skipping questions is smaller between the Control Group and the Negative Treatment (2.155 vs. 1.992, $p = 0.071$) but the share of correct answers is higher in the Negative Treatment (0.5049 vs. 0.5430, $p = 0.035$). These are indications that the increase of correct answers is driven by at least two distinct mechanisms. While loss aversion can explain that pupils take more risky decisions in the Loss Treatment, loss aversion seems not to be induced in the Negative Treatment as the number of omitted answers does not differ from the Control Group. As discussed in Hypothesis 2, pupils instead seem to adjust to the incurred loss of -20 points and seem to be motivated to exert effort due to the increased salience of the "0 point threshold".

Figure 1 shows the average number of omitted questions (left) and the average share of correct answers (right) of pupils by treatments.

Figure 1: Average number of omitted answers and share of correct answers



Note: This figure reports the average number of omitted answers (left) and the average share of correct answers (right) for the Control Group, Loss Treatment and Negative Treatment. Pupils in the Loss Treatment significantly omit more answers than in the Control Group but do not increase the share of correct answers. Pupils in the Negative Treatment do not significantly omit fewer answers but increase the share of correct answers compared to pupils in the Control Group.

Turning to the regression specification confirms the pattern observed in Figure 1. As the data on the number of omitted questions and number of total points show a significant degree of overdispersion (omitted questions: ln α = -0.243 , p-value < 0.001 ; total points: ln α = -2.710, p-value < 0.001), the negative binomial provides a basis for a more efficient estimation for these two outcome variables. For purposes of estimating treatment effects on the share of correct answers, a linear model is applied (OLS).

Table 5 reports on the average treatment effects of the Loss and Negative Treatment on: (1) the number of correct answers, (2) the number of omitted answers, (3) the share of correct answers, and (4) the final points in the test controlling for pupil and class covariates and school fixed effects. In the Loss Treatment, the positive change in correct answers is driven by the fact that pupils skip fewer questions which seems to be driven by an increase in risk taking. Pupils skip significantly fewer questions—respectively answer more questions—than pupils in the Control Group (-0.817, p < 0.001) but do not differ with respect to the share of correct answers. The size of the coefficient for the share of correct answers is close to zero and statistically not significant (0.001, p = 0.963). Interestingly, the share of correct answers in the Control Group is 50.49% and 49.88% in the Loss Treatment. Thus, pupils in the Control Group and Loss Treatment are indifferent between answering or skipping a question but loss framing leads to an increase in answered questions.[44]

Pupils in the Negative Treatment also increase the number of correct answers but, contrary to pupils in the Loss Treatment, do not skip significantly fewer questions than pupils in the Control Group (-0.333, p = 0.106). Nevertheless, the share of correct answers is significantly higher (0.034, p = 0.072).

Although pupils in the Loss and Negative Treatment answer significantly more questions correctly, they do not receive more points in the test. Coefficients for the total points in the test are positive for the Loss Treatment (0.178, p = 0.765) and Negative Treatment (0.846, p = 0.196) but statistically not significant. This is not surprising in the Loss Treatment as the probability to answer a question correctly is roughly 50% and hence the expected value (points) of answering a question is the same as omitting a question. As the

[44]The expected value of answering a question with a success probability of 50% is 2 which equals the value of skipping a question.

probability of a correct answer is similar in the Control Group and in the Loss Treatment, differences in the number of answered and skipped questions should not change the number of total points. Moreover, the insignificant effects on the number of total points in both treatment groups and the insignificant effect on the share of correct answer in the Loss Treatment could be due to a lack of power.

To summarize, pupils in the Loss Treatment answer more questions than pupils in the Control Group but do not increase their accuracy level. In contrast, there is no significant difference in the number of skipped questions between the Negative Treatment and the Control Group. However, pupils in the Negative Treatment increase their level of accuracy.

Result 3 *Pupils in the Loss Treatment answer more questions (take more risky decisions) whereas pupils in the Negative Treatment increase the share of correct answers (answer more accurately).*

Table 5: Treatment Effects - All outcome variables

	(1)	(2)	(3)	(4)
	Correct Answers	*Omitted Answers*	*Share Correct Answers*	*Points in Test*
Treatments				
Loss	0.436***	−0.817***	0.001	0.178
	(0.140)	(0.184)	(0.017)	(0.595)
Negative	0.309**	−0.333	0.034*	0.846
	(0.143)	(0.206)	(0.019)	(0.654)
Controls				
ClassCov	Yes	Yes	Yes	Yes
PupilCov	Yes	Yes	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes
N	1333	1333	1330	1333

Note: This table reports marginal treatment effects on the number of correct answers (1), on the number of omitted items (2), on the share of correct answers (3) and on the number of points in the test (4) including school fixed effects. Covariates: last midterm grade, gender, number of books at home, academic year (grade three or four), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71. Robustness checks with OLS regressions (see Table 15 in the Appendix) and estimation of treatment effects without any controls except including school fixed effects (see Table 12 in the Appendix) show similar results.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6.3 Heterogeneous treatment effects by ability?

In the following, I examine how pupils with different mathematical skill levels respond to the Loss and Negative Treatment and whether heterogeneous gender effects exist.

Ability Based on externally given midterm grades, the effectiveness of framing can be analyzed for different ability levels (low-, middle- and high-ability) which constitutes a novel contribution of this paper. Grades in Germany run from 1+ (excellent) to 6- (insufficient), high-ability pupils refer therefore to those with midterm grades of +1 to 2-; middle-ability pupils have midterm grades of 3+ to 3- and low-ability pupils are those with midterm grades of 4+ to 5-.[45] By asking pupils in the questionnaire about their affinity for mathematics on a 1 (not at all) to 5 (very much) scale, it can be approximated whether low- and high-ability pupils differ in their intrinsic motivation. High-performers have a significantly higher affinity towards mathematics (3.94) than middle- (3.52) and low-performers (3.16).[46] This is an indication that loss-framing might lead to different treatment effects as test score expectations are likely to vary with pupils' ability.

Table 6 reports on the average treatment effects for low-, middle- and high-ability pupils. High-ability pupils are effected positively by both treatments in almost all outcome variables. In the Loss Treatment, high-performers give significantly more correct answers (0.783, $p < 0.001$), skip fewer questions (-0.888, $p < 0.001$) and have higher test scores (1.418, $p = 0.057$) than high-performers in the Control Group. Similar results in size and significance can be found for high-ability pupils in the Negative Treatment [number correct (0.722, $p < 0.001$), number omitted (-0.537, $p = 0.012$), points test (1.974, $p = 0.004$)]. Moreover, the accuracy level also increases significantly (0.057, $p = 0.003$) for high-performers in the Negative Treatment. Differences between high-performers in the Loss and Negative Treatment are not significant except for the number of skipped questions ($p = 0.045$), indicating that the "risk-seeking" effect is larger in the Loss Treatment.

Middle-ability pupils in both treatments do not differ from middle-performers in the Control Group, except that they are significantly more risk-seeking in the Loss Treatment (-0.963, $p = 0.002$) which shows that predictions made based on prospect theory seem to be robust. Differences between the Loss and Negative Treatment are significant for the number of correct answers and the number of omitted answers but overall it seems that middle-performers are not affected by any treatment compared to the Control Group.

Turning to low-ability pupils reveals contrary treatment effects for pupils in the Loss and Negative Treatment. While all coefficients are positive in the Negative Treatment but only significant for the share of correct answers, all coefficients are negative and significant—except for the number of correct answers—in the Loss Treatment. More importantly, all differences between the Loss and Negative Treatment are significant, indicating that the Negative Treatment is superior to the Loss Treatment for low-performers. This could be explained by the fact that low-performers in the Loss Treatment substitute questions which they normally would have skipped by wrong answers. They answer significantly more questions but also increase significantly the number of wrong answer because they might not be able to increase their cognitive performance in the short-run.

The results on ability level do not change if a different grouping of midterm grades is applied. Table 16 in Appendix A.4 presents results for single grouped midterm grades and shows that the positive effects for high-ability pupils is driven by pupils with midterm grades of 2+ to 2-. Coefficients for pupils with midterm grades of 1+ to 1- could be insignificant due to a ceiling effect.[47] Although these pupils are not the highest performers of a class, they still perform good and above average.[48]

[45] In my sample, there was no child with a midterm grade of 6.

[46] The difference between high-ability pupils and middle-ability pupils as well as the difference between middle-ability pupils and low-ability pupils is significant on the 1%-level.

[47] Pupils with a midterm grade of 4 and 5 are grouped because there were in total only 25 pupils with a midterm grade of 5. The groups of *Low-* and *Middle-Ability Pupils* do not change but the group of *High-Ability Pupils* is splitted into midterm grades 1 and midterm grades 2.

[48] Grade 1 is assigned if the performance meets the requirements in an outstanding degree; grade 2 if the performance completely meets the requirements; grade 3 if the performance generally meets the requirements; grade 4 if the performance has shortcomings but as a whole still meets the requirements and grade 5 if the performance does not meet the requirements

To summarize, the Loss and Negative Treatment work similarly well to increase the test performance of high-ability pupils. Nevertheless, the Loss and Negative Treatment have opposite effects on low-ability pupils. Furthermore, Hypothesis 3 cannot be confirmed as the size of treatment effects is not smaller for low-ability pupils. Policy makers should therefore be cautious in implementing loss framing and might prefer the Negative Treatment over the Loss Treatment as performance of low-ability pupils decreases in the latter but not in the Negative Treatment.

Result 4 *The Negative Treatment is superior to the Loss Treatment as performance of low-ability pupils does not decrease. High-ability pupils increase performance in the Negative Treatment and in the Loss Treatment.*

Table 6: Treatment Effects by Ability

	(1) Correct Answers	(2) Omitted Answers	(3) Share Correct Answers	(4) Points in Test
Low-Ability Pupils				
Loss	−0.314	−1.175***	−0.109***	−3.624***
	(0.201)	(0.414)	(0.025)	(0.922)
Negative	0.195	0.584	0.076*	2.150
	(0.350)	(0.750)	(0.044)	(1.473)
N	205	205	205	205
Middle-Ability Pupils				
Loss	0.271	−0.963***	−0.009	−0.717
	(0.197)	(0.318)	(0.025)	(0.850)
Negative	−0.191	−0.240	−0.015	−1.517
	(0.223)	(0.409)	(0.030)	(0.972)
N	376	376	375	376
High-Ability Pupils				
Loss	0.783***	−0.888***	0.026	1.418*
	(0.182)	(0.200)	(0.021)	(0.746)
Negative	0.722***	−0.537**	0.057***	1.974***
	(0.177)	(0.213)	(0.019)	(0.680)
N	755	755	753	755

Note: This table reports average treatment effects of separate regressions for high-, middle-, and low-ability pupils including pupil and class covariates as well as school fixed effects. Covariates:
gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. Standard errors are reported in parentheses and clustered on classroom-level. Robustness checks with OLS regressions show similar results.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

but indicates that the necessary basic knowledge exists and that shortcomings can be resolved in the near future (see https://www.schulministerium.nrw.de/docs/Recht/Schulrecht/Schulgesetz/Schulgesetz.pdf).

Gender The literature has identified gender differences in risk preferences [see Croson and Gneezy, 2009, Eckel and Grossman, 2008, for a review] and Apostolova-Mihaylova et al. [2015] find that loss framing increases on average the final course grade of males but decreases the grade of females. Hence, it is of interest whether heterogeneous gender effects exist also for the Loss and Negative Treatment.

Table 13 in Appendix A.3 presents average treatment effects on all outcome variables separately for boys and girls. In the Loss Treatment, boys (0.413, p = 0.013) as well as girls (0.460, p = 0.014) increase significantly the number of correct answers and also skip significantly fewer questions than boys and girls in the Control Group (boys: -0.867, p < 0.001; girls: -0.752, p = 0.001). In the Negative Treatment, the coefficient for the number of correct answers is positive and significant for girls (0.361, p = 0.083) but not for boys (0.262, p = 0.117). Furthermore, boys and girls in the Negative Treatment tend to skip more questions. This effect is significant for boys but not for girls (boys: -0.373, p = 0.088; girls: -0.284, p = 0.276). Overall, gender differences in all outcome variables are neither significant in the Loss nor in the Negative Treatment.

Interestingly, descriptive statistics suggest that females in the Negative Treatment tend to give the same amount of correct answers and skip an equal amount of questions than boys in the Control Group (see Figure 2 in Appendix B). This is an indication that the Negative Treatment could help to close the gender gap in performance in standardized multiple-choice test which is found in recent studies [see Baldiga, 2014, Pekkarinen, 2015, and the literature mentioned therein]. However, it would need further research to confirm this observation.

The findings on total points in the test (column 4) in the Loss Treatment can be compared to the results of Apostolova-Mihaylova et al. [2015] as the authors focus on the effect of loss framing on students' final course grade. Contrary to Apostolova-Mihaylova et al. [2015], boys in the Loss Treatment score on average 0.183 points lower than boys in the Control Group and females score 0.551 points higher than females in the Control Group. However, neither the coefficients nor the difference between males and females in the Loss Treatment are significant at conventional levels. These opposite findings to Apostolova-Mihaylova et al. [2015] could be driven by pupils' age or the time horizon of the intervention.

Result 5 *There are no detectable heterogeneous gender effects on performance when the grading scheme is manipulated.*

7 Discussion

Here, I want to discuss whether the number of omitted answers can be seen as a risk-taking measure and address three further questions: Do pupils in the Loss Treatment answer marginally more difficult questions? Do pupils change their answering behavior when they reach the threshold of "passing"? Which questions are considered as difficult and do pupils in the Loss Treatment answer strategically by choosing more easy questions?

Number of omitted answers as measure for effort rather than risk-taking? As the test is low stakes, the number of skipped questions can also be viewed as a proxy of effort rather than a measure of risk-seeking, i.e. whether pupils even bother to fill in the questions [Zamarro et al., 2016]. If pupils do not bother to fill in the test questions then the overall share of pupils answering the questions should be rather low, say below 50 % and the share of answered questions should decline for the questions at the end of the test. Table 14 in Appendix A.3 presents descriptive statistics for each test item. *Correct Answer* is the share of pupils—on all pupils giving an answer—who answer the question correctly and *Question Answered*

is the share of pupils who did not skip the question. Overall, pupils seem to be motivated to answer the questions as the share of answered questions is very high and above 70 % in 8 out of 10 questions. The share of pupils answering the question is for example higher in question 9—a question at the end of the test—than in question 1. It seems that less pupils answer the last question (question 10) although the share of pupils giving an answer is still quite high (around 60 %). However, it seems that question 10 is more difficult than other questions as the share of correct answers is much lower compared to the other questions.

Do pupils in the Loss Treatment answer marginally more difficult questions? Pupils in the Loss Treatment were found to not increase the share of correct answers compared to pupils in the Control Group. However, they answer significantly more questions and hence it is conceivable that the marginally answered question is more difficult from an individual point of view. If pupils answer marginally more difficult questions in the Loss Treatment, this should be taken into account in the analysis by e.g. assigning different weights to questions. This, in turn, could then result in a positive and significant treatment effect. To do so, I would need to identify the marginal answered questions for each individual. However, this is not possible due to the pen-and-paper testing format.

Do pupils in the Negative Treatment change their behavior if they reach the threshold of "passing"? On average, pupils in the Negative Treatment increased the number of correct answers compared to pupils in the Control Group. A question of interest is whether and how pupils change their behavior when they accumulated 20 points and hence reached the "passing" threshold. Does performance decline when they reach the positive domain of points? In order to answer this question, I would need to know the exact order of answered questions for each individual. Unfortunately, this is not possible due to the pen-and-paper testing format. Nevertheless, a change in pupils' behavior would be implicit rather than explicit as pupils did not get feedback about their performance during the test. Therefore they could not know how they performed with other questions but they could have formed a belief on whether they are below or above the threshold.

Figure 12 in Appendix B shows Kernel density estimates for the number of points in the test for the Control Group and Negative Treatment. Points for the Negative Treatment have been adjusted to the negative endowment for a better comparison to the Control Group. It seems that fewer pupils in the Negative Treatment score below the threshold of 0 points and that more pupils end up in the top quarter of the points distribution. However, if pupils would have implicitly changed their behavior after passing the threshold, say, a decrease in cognitive effort, a larger share of pupils should be scoring between 20 and 30 points. Thus, either pupils do not know explicitly or implicitly when they reached the threshold, or there is a constant motivational effect of the Negative Treatment. Indications for the latter can be found in Figure 3 in Appendix B. In Figure 3 it is assumed that pupils answered the questions according to the predefined order of questions, question 1 to question 10, and represents Kernel density estimates for the accumulated points in question 5—the first question in which pupils could reach 20 points. It seems that pupils in the Negative Treatment are more motivated to accumulate 20 Points after 5 questions than pupils in the Loss Treatment and Control Group. Figure 4 in Appendix B shows Kernel density estimates of the accumulated number of points at question 10 for pupils who reached 20 points in question 5. Again, it does not seem that pupils change their behavior—decrease performance—after reaching the threshold in the Negative Treatment.

Do pupils in the Negative Treatment answer strategically? Pupils in the Negative Treatment answer the same amount of questions as pupils in the Control Group. However, they answer these questions more

accurately. Hence, the question is whether they answer strategically, say, focus on the 6 out of 10 easiest questions? Do they skip difficult questions to a larger extend than pupils in the Control Group?

Overall, there is no indication that some questions are considered as difficult for pupils in one treatment group but not for pupils in other treatment groups (Table 14 in Appendix A.3). However, questions 3, 6, 8, 9 and 10 seem to be difficult as—across treatment groups—the share of pupils answering these questions correctly is below 50%. Moreover, pupils in the Negative Treatment do not seem to answer some questions more frequently than pupils in the Control Group (*Question Answered*) which is further indication that they do not answer strategically.

8 Conclusion

This paper presents the results of a field experiment in elementary schools in Germany on the effectiveness of loss and gain framing in a mathematical multiple-choice test. Pupils are endowed with the maximum number of points and earning points is framed as a loss in one treatment (Loss Treatment), whereas in another treatment, pupils are endowed with a negative number of points but earning points is framed as a gain (Negative Treatment). These two treatments are then compared to a "traditional" grading scheme in which pupils start with 0 points and earning points is framed as a gain.

The overall finding is that loss framing and a negative downward shift of the point scale have heterogeneous effects on performance (points in test) by ability. In the loss framing condition, high-performers significantly increase their performance while low-performers significantly get lower test scores compared to the Control Group. In contrast, there is no negative treatment effect observable for low-performers in the Negative Treatment but high-performers increase their performance compared to the Control Group.

With regard to the number of correct answers, a measure of performance more interesting for teacher, pupils in both treatment groups answer significantly more questions correctly compared to pupils that are graded "traditionally". These improvements seem to be driven by two different mechanisms. In line with prospect theory [Kahneman and Tversky, 1979], pupils in the Loss Treatment show an increased risk-seeking behavior—increase in answered questions but no decrease in the share of correct answers—whereas pupils in the Negative Treatment answer questions more accurately—the same amount of answered questions but an increase in the share of correct answers.[49]

Although the experimental design has some limitations—treatment effects can only be interpreted for the populations studied; short-run and low-stakes intervention—the results give valuable insights to educators and policy-makers who aim to apply insights from behavioral economics into their fields. While loss framing might seem appealing to implement in the educational system as it represents a promising and cost-effective intervention, my results show that low-performers—which are usually the target audience of policy interventions—are made worse off. Moreover, the experimental design allows the isolation of the effort effect from the learning effect, showing that differences in performance in the Negative Treatment are likely to be driven by an increase in cognitive effort. This insight is interesting as it shows that success is not based solely on innate ability. Hence, it might be effective to teach pupils that exerting effort while taking a test is as important as motivating pupils to put effort into learning.

[49] The finding of increased risk-seeking behavior persists if pupils are differentiated by gender or ability level.

While there are a number of laboratory and some field studies exploiting the effectiveness of loss framing [Hossain and List, 2012, Apostolova-Mihaylova et al., 2015, Fryer et al., 2012], there are only a few field experiments applying loss framing in an educational setting and only one study in elementary and high schools [Levitt et al., 2016, Apostolova-Mihaylova et al., 2015, Fryer et al., 2012]. My study is one of the first studies showing how framing manipulations change the behavior for pupils of different ability levels and sheds light on the underlying mechanism. Furthermore, my results suggest that—besides loss framing—there are further promising and cost-effective methods to boost performance, for example, a downward shift of the point scale. However, it remains for future research to analyze the impact of framing effects in high-stakes testing environments and in long-run interventions to get a more comprehensive picture of behavioral interventions in the educational sector and the workplace.

References

Richard Allington, Anne McGill-Franzen, Gregory Camilli, Lunetta Williams, Jennifer Graff, Jacqueline Zeig, Courtney Zmach, and Rhonda Nowak. Addressing Summer Reading Setback Among Economically Disadvantaged Elementary Students. *Reading Psychology*, 31(5):411–427, 2010.

Yvonne Anders, Nele McElvany, and Jürgen Baumert. Die Einschätzung lernrelevanter Schülermerkmale zum Zeitpunkt des Übergangs von der Grundschule auf die weiterführende Schule. Wie differenziert urteilen Lehrkräfte? In Kai Maaz, Jürgen Baumert, Cornelia Gresch, and Nele McElvany, editors, *Der Übergang von der Grundschule in die weiterführende Schule. Leistungsgerechtigkeit und regionale, soziale und ethnisch–kulturelle Disparitäten*, Bildungsforschung. 34, pages 313–330. Bundesministerium für Bildung und Forschung, Referat Bildungsforschung, Bonn, 2010.

Joshua Angrist. Conditional Independence in Sample Selection Models. *Economics Letters*, 54(2):103–112, 1997.

Maria Apostolova-Mihaylova, William Cooper, Gail Hoyt, and Emily Marshall. Heterogeneous Gender Effects under Loss Aversion in the Economics Classroom: A Field Experiment. *Southern Economic Journal*, 81 (4):980–994, 2015.

Olivier Armantier and Amadou Boly. Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada. *The Economic Journal*, 123(573):1168–1187, 2013.

Olivier Armantier and Amadou Boly. Framing of Incentives and Effort Provision. *International Economic Review*, 56(3):917–938, 2015.

Nava Ashraf, Oriana Bandiera, and Scott Lee. Awards Unbundled: Evidence from a Natural Field Experiment. *Journal of Economic Behavior & Organization*, 100:44 – 63, 2014.

Katherine Baldiga. Gender Differences in Willingness to Guess. *Management Science*, 60(2):434–448, 2014.

Jere Behrman, Susan Parker, Petra Todd, and Kenneth Wolpin. Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy*, 123(2):325–364, 2015.

Roland Bénabou and Jean Tirole. Incentives and Prosocial Behavior. *The American Economic Review*, 96 (5):1652–1678, 2006.

Christiane Bradler and Susanne Neckermann. The Magic of the Personal Touch: Field Experimental Evidence on Money and Appreciation as Gifts. Discussion Paper 16-045/VII, Tinbergen Institute, May 2016.

Christiane Bradler, Robert Dur, Susanne Neckermann, and Arjan Non. Employee Recognition and Performance: A Field Experiment. *Management Science*, accepted, 2016.

Miriam Bruhn and David McKenzie. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, 1(4):200–232, 2009.

Leonardo Bursztyn and Robert Jensen. How Does Peer Pressure Affect Educational Investments? *The Quarterly Journal of Economics*, 130(3):1329–1367, 2015.

David Card. Chapter 30 - The Causal Effect of Education on Earnings. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, Part A, pages 1801–1863. Elsevier, 1999.

David Card and Alan Krueger. Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States. *Journal of Political Economy*, 100(1):1–40, February 1992.

Jennifer Henderlong Corpus, Megan McClintic-Gilbert, and Amynta Hayenga. Within-Year Changes in Children's Intrinsic and Extrinsic Motivational Orientations: Contextual Predictors and Academic Outcomes. *Contemporary Educational Psychology*, 34(2):154–166, 2009.

Rachel Croson and Uri Gneezy. Gender Differences in Preferences. *Journal of Economic Literature*, 47(2): 448–474, 2009.

Flavio Cunha and James Heckman. The Technology of Skill Formation. *American Economic Review*, 97(2): 31–47, 2007.

Eszter Czibor, Sander Onderstal, Randolph Sloof, and Mirjam Van Praag. Does Relative Grading Help Male Students? Evidence from a Field Experiment in the Classroom. Discussion Paper 14-116/V, Tinbergen Institute, 2014.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Using Randomization in Development Economics Research: A Toolkit. *Handbook of Development Economics*, 4:3895–3962, 2007.

Christian Dustmann, Patrick Puhani, and Uta Schönberg. The Long–Term Effects of Early Track Choice. *Economic Journal*, accepted, 2016.

Catherine Eckel and Philip Grossman. Chapter 113 - Men, Women and Risk Aversion: Experimental Evidence. In Charles Plott and Vernon Smith, editors, *Handbook of Experimental Economics Results*, volume 1, pages 1061–1073. Elsevier, 2008.

Roland Fryer. Financial Incentives and Student Achievement: Evidence from Randomized Trials. *The Quarterly Journal of Economics*, 126(4):1755–1798, 2011a.

Roland Fryer. Financial Incentives and Student Achievement: Evidence from Randomized Trials. *The Quarterly Journal of Economics*, 126(4):1755–1798, 2011b.

Roland Fryer, Steven Levitt, John List, and Sally Sadoff. Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. Working Paper 18237, National Bureau of Economic Research, July 2012.

Thomas Fuchs and Ludger Wößmann. What Accounts for International Differences in Student Performance? A Re–examination Using PISA Data. *Empirical Economics*, 32(2-3):433–464, 2007.

Uri Gneezy and Aldo Rustichini. Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics*, 115(3):791–810, 2000.

Sebastian Goerg and Sebastian Kube. Goals (th)at Work–Goals, Monetary Incentives, and Workers' Performance. *MPI Collective Goods Preprint*, (2012/19), 2012.

Eric Hanushek, Guido Schwerdt, Simon Wiederhold, and Ludger Wößmann. Returns to skills around the world: Evidence from PIAAC. *European Economic Review*, 73:103–130, 2015.

Douglas Harris and Tim Sass. Teacher Training, Teacher Quality and Student Achievement. *Journal of Public Economics*, 95(7):798–812, 2011.

Fuhai Hong, Tanjim Hossain, and John List. Framing Manipulations in Contests: A Natural Field Experiment. *Journal of Economic Behavior & Organization*, 118:372 – 382, 2015.

Tanjim Hossain and John List. The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science*, 58(12):2151–2167, 2012.

Nina Jalava, Juanna Schrøter Joensen, and Elin Pellas. Grades and Rank: Impacts of Non-Financial Incentives on Test Performance. *Journal of Economic Behavior & Organization*, 115:161–196, 2015.

Lene Arnett Jensen, Jeffrey Jensen Arnett, Shirley Feldman, and Elizabeth Cauffman. It's Wrong, but Everybody Does It: Academic Dishonesty among High School and College Students. *Contemporary Educational Psychology*, 27(2):209–228, 2002.

Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–91, 1979.

James S. Kim. The effects of a voluntary summer reading intervention on reading activities and reading achievement. *Journal of Educational Psychology*, 99(3):505–515, 2007. URL http://psycnet.apa.org/journals/edu/99/3/505/.

Alexander Koch, Julia Nafziger, and Helena Skyt Nielsen. Behavioral Economics of Education. *Journal of Economic Behavior & Organization*, 115:3–17, 2015.

Michal Krawczyk. Framing in the Field: A Simple Experiment on the Reflection Effect. Technical Report 14, University of Warsaw, 2011.

Sebastian Kube, Michel André Maréchal, and Clemens Puppea. The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 102(4):1644–62, June 2012.

Adam Lavecchia, Heidi Liu, and Philip Oreopoulos. Chapter 1 - Behavioral Economics of Education: Progress and Possibilities. In Stephen Machin Eric Hanushek and Ludger Wößmann, editors, *Handbook of the Economics of Education*, volume 5, pages 1 – 74. Elsevier, 2016.

Steven Levitt, John List, Susanne Neckermann, and Sally Sadoff. The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. Discussion Paper 12-38, Centre for European Economic Research (ZEW), 2012.

Steven Levitt, John List, Susanne Neckermann, and Sally Sadoff. The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *American Economic Journal: Economic Policy*, 8(4):183–219, 2016.

John List, Azeem Shaikh, and Yang Xu. Multiple Hypothesis Testing in Experimental Economics. Working Paper 21875, National Bureau of Economic Research, January 2016.

David McEvoy. Loss Aversion and Student Achievement. *Economics Bulletin*, 36(3):1762–1770, 2016.

Steffen Mueller. Teacher Experience and the Class Size Effect – Experimental Evidence. *Journal of Public Economics*, 98(0):44–52, 2013.

Susanne Neckermann, Reto Cueni, and Bruno Frey. Awards at Work. *Labour Economics*, 31:205 – 217, 2014.

Frank Pajares and Laura Graham. Self-Efficacy, Motivation Constructs, and Mathematics Performance of Entering Middle School Students. *Contemporary Educational Psychology*, 24(2):124–139, 1999.

Tuomas Pekkarinen. Gender Differences in Behaviour under Competitive Pressure: Evidence on Omission Patterns in University Entrance Examinations. *Journal of Economic Behavior & Organization*, 115:94 – 110, 2015.

Sally Sadoff. The Role of Experimentation in Education Policy. *Oxford Review of Economic Policy*, 30(4): 597–620, 2014.

Matthew Springer, Brooks Rosenquist, and Walker Swain. Monetary and Nonmonetary Student Incentives for Tutoring Services: A Randomized Controlled Trial. *Journal of Research on Educational Effectiveness*, 8(4):453–474, 2015.

Richard Thaler and Eric Johnson. Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*, 36(6):643–660, 1990. doi: 10.1287/mnsc.36.6.643.

Valentin Wagner and Gerhard Riener. Peers or Parents? On Non-Monetary Incentives in Schools. Discussion paper, Düsseldorf Institute for Competition Economics (DICE), November 2015.

Ludger Wößmann. The Effect Heterogeneity of Central Examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, 13(2):143–169, 2005.

Gema Zamarro, Collin Hitt, and Ildefonso Mendez. When Students Don't Care: Reexamining International Differences in Achievement and Non-Cognitive Skills. (18), October 2016.

A Tables

A.1 Randomization Table

Table 7: Sample Size by Gender and Treatment

	Control	*Loss*	*Negative*	*Overall*
Full Sample				
N individuals	515	468	394	1377
Correct Answers	3.915	4.165	4.246	4.094
	(2.173)	(2.239)	(2.344)	(2.248)
Points Test	19.695	19.876	20.995	20.229
	(8.105)	(8.255)	(8.458)	(8.266)
Boys				
N individuals	254	227	203	684
Correct Answers	4.201	4.436	4.379	4.332
	(2.220)	(2.198)	(2.384)	(2.262)
Points Test	20.661	20.326	21.182	20.705
	(8.201)	(8.301)	(8.689)	(8.376)
Girls				
N individuals	246	224	182	652
Correct Answers	3.650	3.951	4.176	3.900
	(2.092)	(2.277)	(2.294)	(2.221)
Points Test	19.187	19.473	20.857	19.752
	(8.062)	(8.398)	(8.352)	(8.277)
Numb. Classes	26	23	21	71

Note: The table displays the descriptive statistics (means) of the number of pupils, number of correct answers and test scores in each of the treatment groups and the control group. 20 points have been added to the Negative Treatment to adjust for the negative endowment. Standard deviations are displayed in parentheses. In my final analysis, 1.333 observations are included. 41 pupils did not report their gender.

Table 8: Randomization Check

(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Treatments	**DI**	**p-values**			
			Unadj.	Multiplicity Adj.		
			Remark 3.1	Thm. 3.1	Bonf.	Holm
Age	Control vs. Loss	0.0593	0.2227	0.9147	1.0000	1.0000
	Control vs. Negative	0.0819	0.1217	0.8023	1.0000	1.0000
Month of Birth	Control vs. Loss	0.0831	0.7140	0.9793	1.0000	1.0000
	Control vs. Negative	0.1552	0.5087	0.9813	1.0000	1.0000
Num. Older Sib.	Control vs. Loss	0.0055	0.9307	0.9307	1.0000	0.9307
	Control vs. Negative	0.1043	0.1473	0.8473	1.0000	1.0000
Female Pupil	Control vs. Loss	0.0047	0.8800	0.9840	1.0000	1.0000
	Control vs. Negative	0.0193	0.5883	0.9697	1.0000	1.0000
Language German	Control vs. Loss	0.0699	0.0547**	0.5453	0.8747	0.8200
	Control vs. Negative	0.0351	0.3203	0.9500	1.0000	1.0000
Remedial Teaching	Control vs. Loss	0.0229	0.1593	0.8467	1.0000	1.0000
	Control vs. Negative	0.0227	0.0990*	0.7403	1.0000	1.0000
Teacher Exp.	Control vs. Loss	0.4606	0.5047	0.9910	1.0000	1.000
	Control vs. Negative	4.0972	0.0003***	0.0003***	0.0053***	0.0053***
Unemployment	Control vs. Loss	0.0017	0.5797	0.9877	1.0000	1.0000
	Control vs. Negative	0.0033	0.2810	0.9387	1.0000	1.0000
Books Home	Control vs. Loss	0.101	0.3627	0.7473	1.0000	1.0000
	Control vs. Negative	0.245	0.0337**	0.1213	0.2693	0.1347
Summer Holiday	Control vs. Loss	0.022	0.4650	0.7100	1.0000	0.9300
	Control vs. Negative	0.095	0.0033***	0.0183**	0.0267**	0.0200**
Day diff. Holiday	Control vs. Loss	5.536	0.0003***	0.0003***	0.0027***	0.0027***
	Control vs. Negative	6.252	0.0003***	0.0003***	0.0027***	0.0023***
Academic Year	Control vs. Loss	0.058	0.0337**	0.1477	0.2693	0.1683
	Control vs. Negative	0.007	0.7953	0.7953	1.0000	0.7953

Note: This table presents randomization checks for control variables used in the analysis adjusting for multiple hypothesis testing. *DI* is the difference in means between the Control Group and each of the treatment groups. Columns 4-7 display p-values. Column (4) presents multiplicity-unadjusted p-value; columns (5)-(7) display multiplicity-adjusted p-values. See also List et al. [2016] on multiple hypothesis testing. Differences in baseline scores and concerns about non-random self-selection are discussed in subsection 5.1.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2 Self-selection

Table 9: Self-selection by Treatment

	Control Group	*Loss Treatment*	*Negative Treatment*
# absent pupils	4.27	4.13	6.27
% absent pupils	17.71	17.18	25.79
Midterm Grade	6.49	6.68	6.26
N (# classes)	26	23	22

Note: This table reports on the number of pupils absent on the test day and pupils' last midterm grade. Cell entries represent averages on class level. Midterm Grade is measured on a 1 to 15 scale where 1 is the best grade and 15 the worst. In US equivalents a midterm grade of 6 is a B- and 7 a C+. Differences between Control and Treatment Groups are statistically not significant using a simple t-test.

A.3 Estimation Tables

Table 10: Treatment Effects - Number of Omitted Items

	(1)	(2)	(3)	(4)
Treatments				
Loss	-0.760^{***}	-0.787^{***}	-0.832^{***}	-0.817^{***}
	(0.210)	(0.198)	(0.189)	(0.184)
Negative	-0.281	-0.309	-0.286	-0.333
	(0.221)	(0.219)	(0.209)	(0.206)
Controls				
ClassCov	No	Yes	No	Yes
PupilCov	No	No	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes
N	1333	1333	1333	1333

Note: This table reports the marginal effects of a negative binomial regression including school fixed effects. Dependent variable: number of omitted questions. Covariates: last midterm grade, gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. 44 observations are dropped due to missing values. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Treatment Effects - Share of Correct Answers

	(1)	(2)	(3)	(4)
Treatments				
Loss	−0.007	−0.009	0.007	0.001
	(0.021)	(0.020)	(0.018)	(0.017)
Negative	0.054**	0.052**	0.035	0.034*
	(0.025)	(0.023)	(0.023)	(0.019)
Controls				
ClassCov	No	Yes	No	Yes
PupilCov	No	No	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes
N	1330	1330	1330	1330

Note: This table reports the results of a generalized linear model school fixed effects. Dependent variable: share of correct answers ($\frac{\#Correct}{10-\#Omitted}$). Covariates: last midterm grade, gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. 44 observations are dropped due to missing values. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71.
* p < 0.10, ** p < 0.05, *** p < 0.01

Table 12: Treatment Effects without control variables- Correct, Omitted, Share and Points

	(1)	(2)	(3)	(4)
	Correct Answers	*Omitted Answers*	*Share Correct Answers*	*Points in Test*
Treatments				
Loss	0.320	−0.768***	−0.008	−0.053
	(0.213)	(0.211)	(0.020)	(0.704)
Negative	0.482**	−0.271	0.054**	1.584*
	(0.233)	(0.219)	(0.024)	(0.836)
Controls				
ClassCov	No	No	No	No
PupilCov	No	No	No	No
SchoolFE	Yes	Yes	Yes	Yes
N	1377	1377	1374	1377

Note: This table reports marginal treatment effects on the number of correct answers (1), on the number of omitted items (2), on the share of correct answers (3) and on the number of points in the test (4) including only school fixed effects. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Treatment Effects by Gender

Panel A: Regression	(1) *Correct Answers*	(2) *Omitted Answers*	(3) *Share Correct Answers*	(4) *Points in Test*
Treatments				
Loss	0.413**	−0.867***	−0.002	−0.183
	(0.166)	(0.215)	(0.021)	(0.768)
Negative	0.262	−0.373*	0.034	0.552
	(0.167)	(0.219)	(0.021)	(0.779)
Female	−0.248	0.299*	−0.001	−0.379
	(0.165)	(0.174)	(0.021)	(0.677)
Loss × Female	0.047	0.115	0.006	0.734
	(0.211)	(0.259)	(0.027)	(0.942)
Negative × Female	0.099	0.089	0.002	0.600
	(0.245)	(0.251)	(0.030)	(0.970)
Controls				
ClassCov	Yes	Yes	Yes	Yes
PupilCov	Yes	Yes	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes
Panel B: Contrast	*Treatment vs. No Treatment for Females*			
Loss	0.460**	−0.752***	0.004	0.551
	(0.186)	(0.231)	(0.022)	(0.751)
Negative	0.361*	−0.284	0.035	1.152
	(0.208)	(0.260)	(0.027)	(0.846)
N	1333	1333	1330	1333

Note: Panel A reports average treatment effects for boys including school fixed effects; panel B presents average treatment effects for girls. Covariates: last midterm grade, gender, number of books at home, academic year (grade three or four), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71. Robustness checks with OLS regressions show similar results. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Share of correct and answered questions by test item

	Control	*Loss*	*Negative*
Question 1			
Correct Answers	78.63	77.17	80.20
Question Answered	73.59	81.41	76.90
Question 2			
Correct Answers	59.38	55.43	62.92
Question Answered	87.96	92.52	90.36
Question 3			
Correct Answers	36.57	37.91	42.53
Question Answered	75.92	83.97	78.17
Question 4			
Correct Answers	54.59	50.62	55.38
Question Answered	80.39	86.11	82.49
Question 5			
Correct Answers	64.90	67.26	69.27
Question Answered	95.15	95.94	94.16
Question 6			
Correct Answers	37.75	34.94	38.11
Question Answered	87.96	88.68	83.25
Question 7			
Correct Answers	58.10	61.63	63.19
Question Answered	83.88	86.32	82.74
Question 8			
Correct Answers	41.61	46.88	48.50
Question Answered	60.19	68.38	67.51
Question 9			
Correct Answers	39.42	40.40	39.10
Question Answered	79.81	85.68	79.19
Question 10			
Correct Answers	15.91	16.16	21.96
Question Answered	59.81	70.09	64.72

Note: This table reports on the number of correct questions and answered questions separately for each test item. *Correct Answer* is the share of pupils on all pupils giving an answer who answer the question correctly. *Question Answered* is the share of pupils who did not omit the question. Cell entries present percentages.

A.4 Robustness Checks

Table 15: Robustness Check - Correct Answers, Omitted Answers, Points in Test

	Correct Answers		*Omitted Answers*		*Points in Test*	
	OLS	Poisson	OLS	NBREG	OLS	NBREG
Treatments						
Loss	0.452***	0.436***	−0.761***	−0.817***	0.309	0.178
	(0.139)	(0.140)	(0.175)	(0.184)	(0.580)	(0.595)
Negative	0.352**	0.309**	−0.258	−0.333	0.932	0.846
	(0.137)	(0.143)	(0.202)	(0.206)	(0.609)	(0.654)
Controls						
ClassCov	Yes	Yes	Yes	Yes	Yes	Yes
PupilCov	Yes	Yes	Yes	Yes	Yes	Yes
SchoolFE	Yes	Yes	Yes	Yes	Yes	Yes
N	1333	1333	1333	1333	1333	1333

Note: This table compares the results of a linear (OLS) and a negative binomial regression (marginal effects) for the number of correct answers, number of omitted answers and the total points in the test. Covariates: gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

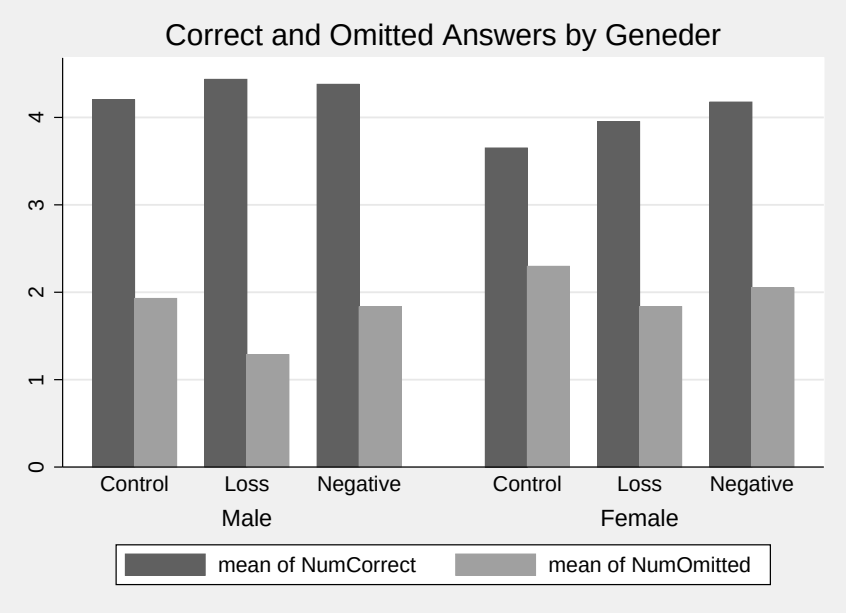
Table 16: Treatment Effects by Midterm Grade

	(1) Correct Answers	(2) Omitted Answers	(3) Share Correct Answers	(4) Points in Test
Midterm Grade = 4+ to 5-				
Loss	−0.314	−1.175***	−0.109***	−3.624***
	(0.201)	(0.414)	(0.025)	(0.922)
Negative	0.195	0.584	0.076*	2.150
	(0.350)	(0.750)	(0.044)	(1.473)
N	205	205	205	205
Midterm Grade = 3+ to 3-				
Loss	0.271	−0.963***	−0.009	−0.717
	(0.197)	(0.318)	(0.025)	(0.850)
Negative	−0.191	−0.240	−0.015	−1.517
	(0.223)	(0.409)	(0.030)	(0.972)
N	376	376	375	376
Midterm Grade = 2+ to 2-				
Loss	0.822***	−0.952***	0.039*	1.641**
	(0.203)	(0.244)	(0.023)	(0.798)
Negative	0.654***	−0.519**	0.060***	1.794***
	(0.176)	(0.254)	(0.021)	(0.689)
N	564	564	562	564
Midterm Grade = 1+ to 1-				
Loss	0.482	−0.448	−0.002	0.832
	(0.342)	(0.282)	(0.036)	(1.218)
Negative	0.567	−0.468**	0.022	1.413
	(0.403)	(0.247)	(0.033)	(1.240)
N	191	191	191	191

Note: This table reports average treatment effects of separate regressions for midterm grades including pupil and class covariates as well as school fixed effects. In comparison to Table 6 in Section 6.3, the group of pupils with a midterm grade of 3+ to 3- (4+ to 5-) is equivalent to the group of *middle-ability pupils* (*low-ability pupils*). In contrast to Section 6.3, the group of *high-ability pupils* is splitted into midterm grades 1+ to 1- and 2+ to 2-. Covariates: gender, number of books at home, academic year (grade 3 or 4), teachers' working experience (in years), day differences between test and next holidays and a dummy whether the test was written before or after the summer break. Standard errors are reported in parentheses and clustered on classroom-level. Robustness checks with OLS regressions show similar results.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

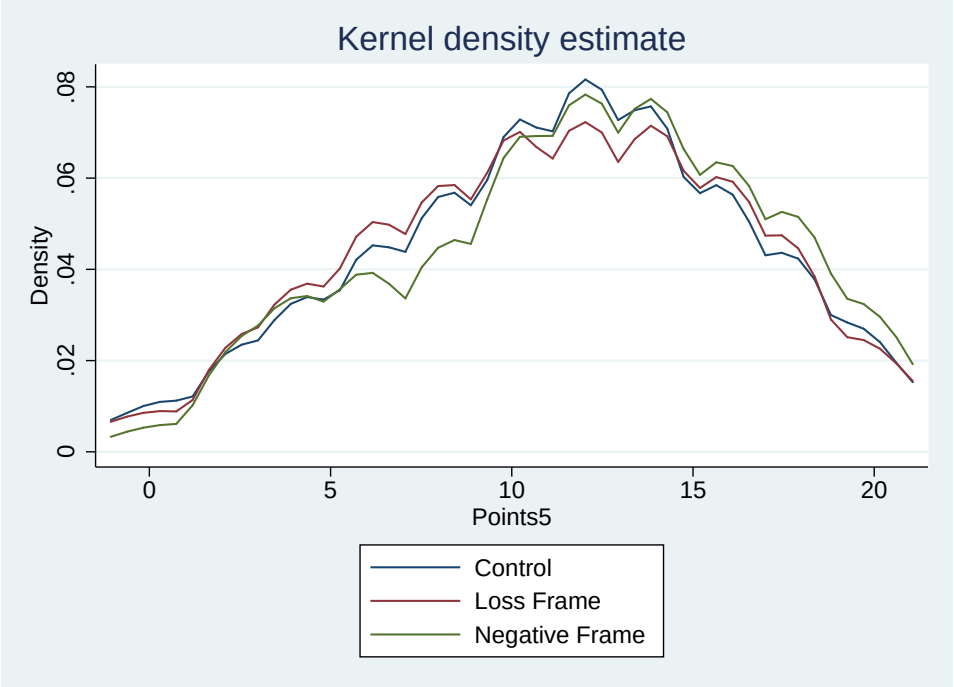
B Figures

Figure 2: Average number of omitted answers and share of correct answers



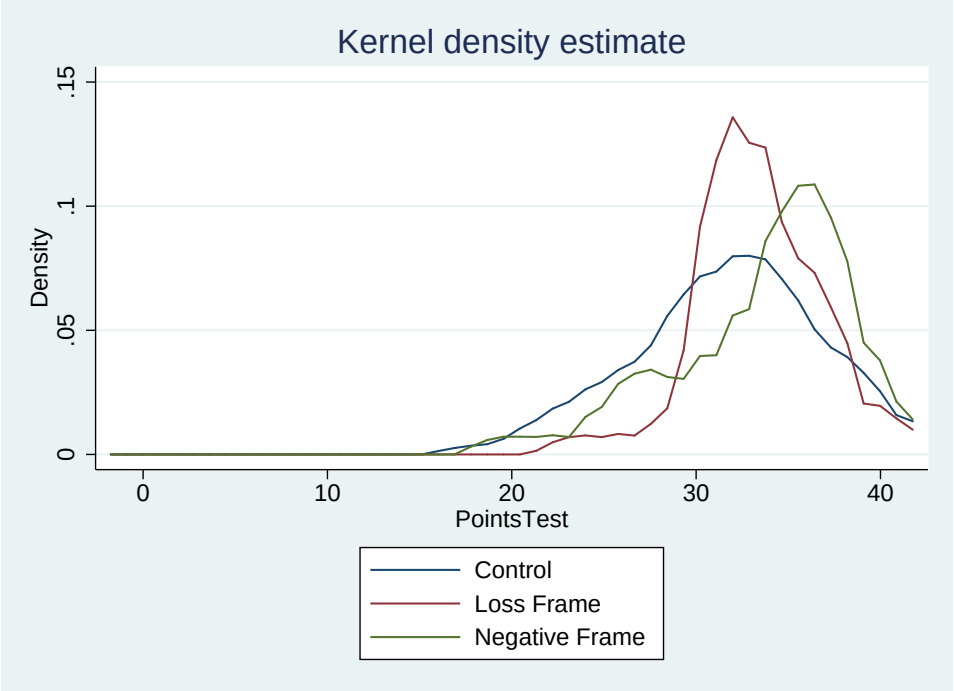
Note: This figure reports the average number of correct and omitted answers separately for boys and girls.

Figure 3: Kernel density plot: Points after five questions (Q1-Q5)



Note: This Figure presents Kernel density estimates for the number of points reached in the first five questions for the Control Group, the Loss Treatment and the Negative Treatment.

Figure 4: Kernel density plot: Final points of pupils who accumulated 20 points at Q5



Note: This Figure presents Kernel density estimates for the number of final points reached in the test for pupils who accumulated 20 points in the first five questions.

C Instructions, Questionnaire and Consent Form

C.1 Instruction for Teacher

—not intended for publication—

The following instructions were given to teachers in the Loss Treatment. Instructions for the Control Group and Negative Treatment contained the same information but the way points could be earned differed as explained in Section 3.

Figure 5: Teacher Instructions—First Letter [translated from German]

Instructions for [class] of [name of school]

Thank you for supporting my research project. Today, I am sending you the instructions for running the test. It is absolutely necessary that the procedure is carried out in the described way to be able to successfully evaluate this project. Otherwise, the experiment cannot be carried out properly and the results are no longer of use. Therefore, you are requested to act according to the instructions given in this letter.

The mathematical test shall be written **until 13.11.2015**. When exactly is up to you. Please choose a testing week in which no other exam is written so that pupils' workload is minimized. In total, you receive two envelopes containing materials to carry out the experiment. In this envelope I have send you instructions on how to announce the test, the preparation material for pupils as well as the consent forms to be signed by parents. In the second envelope you will get further instructions on how exactly to execute the test at the testing day, the actual tests as well as pupil questionnaires. This second envelope is mailed to you close to the testing day. Therefore, it is important that you send me the exact testing date to wagner@dice.hhu.de as soon as you now when the test shall be written.

The test is similar to the Känguru-Wettbewerb. However, the scoring is slightly different from the original test. Pupils in your class start the test with the maximum number of points (40 points). 0 points are deducted for each correct answer, -2 points are deducted for a skipped answers and -4 points are deducted if the answer is wrong. The highest achievable score is 40, the lowest 0. The test takes 30 minutes, is evaluated by us and pupils will receive a score at the end. It is up to you whether you want to assign a grade for the score at the very end.

Test announcement

1. The test will be announced exactly **one week** in advance. Please write the test date on the board. Pupils shall have the opportunity to prepare for the test.

2. Please explain that the test is mandatory and that it will be corrected and evaluated but that it will not count for the report marks. Do not yet explain in which way points are allocated in the test. This will be done immediately before the test on the test day.

3. Please distribute the preparation material thereafter and answer all remaining questions.You can justify the test by saying that you want to try out a different kind of testing format. Otherwise, you could also justify the test by saying that you want to find out in which areas of mathematics pupils need to catch up in the course material. Please refrain from actively motivating pupils to study for the test during this week. Questions about the learning materials or the process of the test can be answered, of course. I also ask you not to tell the pupils that this test is taking place as part of a broader study by the University of Düsseldorf. Please do not mention that other classes also participate in this project.

Please send us an e-mail with the date of the test **on the same day** of announcement. Please do not tell pupils the background of this research project before the actual test was written. Please be not surprised if the test instructions are different for the classes of your colleagues. This is intentional and is part of the research project.

Please contact us by phone or email in case you have any question.

Figure 6: Teacher Instructions—Second Letter [translated from German]

Instructions for the Control Group and Negative Treatment differ in point 2 where the respective allocation of points is explained.

Instructions for [class] of [name of school]
In this envelope you have received the tests, the questionnaires for pupils, a list to enter the midterm grades and a statement of privacy. Please read the instructions carefully and execute the test in the given order:

Execution of the test: time 30 minutes

1. Please let pupils—similar to exams—set the tables a little bit apart. Additionally let them put up a privacy screen between each other. Remind pupils that all questions have to be answered independently and that each attempt to copy from their neighbor will be punished with the removal of the test. If the latter happens, please indicate this by an "X" in the upper right corner of the first page of the test.
2. Before the test starts, please read out aloud the following text to the class: "The test contains a total of 10 tasks that must be solved within 30 minutes. For each task, there are 4 wrong and 1 correct answers. Every one of you starts with the full score, which is 40 points. For each correct answer you get 0 points and for each wrong answer 4 points are deducted. 2 points are deducted if you skip an answer. Calculators are not allowed, but "scratch paper" for sketches and small calculations are allowed, of course!"
3. Please tell pupils that they should not write their names on the test. For privacy reasons, each test already received a "Test-ID number".
4. Now the test starts and lasts 30 minutes in total.
5. While the test is ongoing, please write down the corresponding name for each Test-ID number (upper left corner on the first page of the test) on a sheet of paper. For this, you could also use a class list. This sheet serves as an "encryption key" which you do not send back to us and keep for yourself. This is important so that you know which test belongs to which pupil after you receive the corrected tests from us.
6. After the test, the questionnaires have to be answered. These have already been attached to the test. Again, this is to be filled out independently and quietly.

Please send the tests, questionnaires, preparation sheets and the list with the midterm grades back to us with the enclosed envelope on the same day. The tests are then corrected immediately and sent back to you. Please fill in the midterm grades in the list we have send you. The Test-ID numbers serve here as an encryption key. Example: "Andrea Albers", has the Test-ID number 12, then please write down under the number 12 in the list the midterm grade plus tendency of Andrea Albers. By this method, we can meet the requirements of privacy policy since so it cannot be identified which grade belongs to which pupil retrospectively. In addition, all materials which are handed out during the project will be returned to you. Once all participating schools have conducted the tests, we start with the statistical analysis and send you the results.
Thank you very much.

C.2 Teacher and Student Questionnaire

—not intended for publication—

Figure 7: Teacher Questionnaire [translated from German]

Teacher Questionnaire

Please answer all of the following questions truthfully. The questions are very important for us to gain insights from the teacher perspective. Please send the questionnaire backt to us. A stamped envelope is attached.

School: _____ Class: _____

For how long are you working as a teacher?: _____ Date of test: _____

How many students are in your class? _____ …attend the school (approx.)?: _____

1.In which school hour did you write the test? _____

2. In you oppinion, how difficult is the test for pupils?

1 □	2 □	3 □	4 □	5 □
too easy		medium		too difficult

3. Does your school apply multi-grade teaching? If yes, which grades are teached together?

_____

4. Does your school have media facilities where pupils can acquire media skills?
Yes □ No □

5. If yes, do you actively teach media competencies in your courses?
Yes □ No □

6. Do you plan to participate in a mathematics competition this year (Känguru, Pangea etc.)?
Yes □ No □

7. Did you actively prepare pupils for the test?
Yes □ No □

If yes, how exactly: _____

8. Please rank the social environment of the school district?

1 □	2 □	3 □	4 □	5 □
socially troubled area				Very good residential area

9. Did you inform parents about the study?
Yes □ No □

If yes: before the test □ after the test □

10. On which basis are pupils sorted into classes?

_____

11. Please give us a short feedback on the back.Did you notice anything that could be of relevance for our analysis? Do you have any comments / suggestions for improvements ?

Thank you

Figure 8: Student Questionnaire [translated from German]

Student Questionnaire

Please answer all of the following questions and tick the appropriate boxes. It is very important that you answer all questions truthfully. Your answers will be treated anonymously and no other students in your class will have access to them.

Test-ID: Class: ____

School: _____ Age: _____

Gender: □ Girl □ Boy

Mother tongue: □ German □ other

1. How difficult was the test?: _____

1 □	2 □	3 □	4 □	5 □
too easy		medium		too hard

2. How much do you like the subject mathematics?

□	□	□	□	□
not at all		medium		very much

3. Did you learn for the test?

□ Yes □ No

If yes,

a)How many hours did you approx. learn? _____

b) How many preparation sheets did you solve? _____

4. How many books do you have at home?
Approximately 40 books fit on a meter of bookcase. Please do not count in newspapers and your textbooks.

0-10 □ 11-25 □ 26-100 □ 101–200 □ 201–500 □ more than 500 □

5. How many siblings do you have?:

0 □ 1 □ 2 □ 3 □ more than 3 □

6. How many of your siblings are older than you?

_____

7. In which month is your birthday?

_____

Thank you

C.3 Consent Form

—not intended for publication—

Figure 9: Consent Form to be signed by parents (translated from German)

Dear Parents,
I am a doctoral student of economics at the Heinrich-Heine-University of Düsseldorf and conduct research in the field of empirical economics of education. As part of my thesis, I am currently working on the research project "Motivation in schools".
In this context, I am running a scientific study which will take part from **May to November 2015**. The aim of the study is to analyze pupils' motivation in a mathematical multiple-choice test. Some pupils will start the test with the maximum number of points while others start, as usually, with 0 Points. I then analyze how the initial endowment affects pupils' motivation to exert effort in the test.
The mathematical questions are a compilation of old test questions of the *Känguru-Test* (http://www.mathe-kaenguru.de/). This is a nationwide test with about 886.000 participants last year and which has been conducted for more than over 20 years by the Department of Mathematics of the Humboldt University Berlin. The questions of the *Känguru-Test* are designed in a way that the joy of (mathematical) thinking and working shall be awakened and supported.
I would be delighted if your child would be allowed to participate in the test which will take place in a regular scheduled lesson. For this I need your consent. Please sign the attached consent form and hand it to your child. The teacher will then collect the forms.
Thank you for your cooperation!
Sincerely yours,

Declaration of Consent for study participation

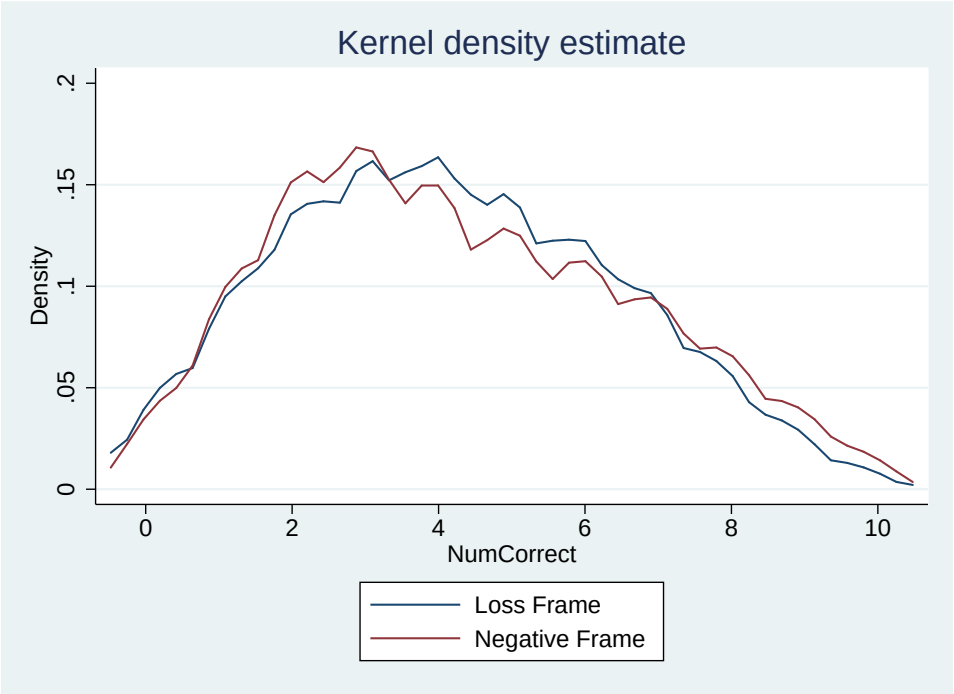
Hereby I (name of parent) voluntarily agree that my child (name of child) born on (date of birth) participates in the project described above and writes the test as part of a lesson. I give my consent that relevant scientific data will be stored and analyzed. My child' data are treated privately and anonymously, so that it is impossible to trace back on my child. It is—for me and my child—always possible to cancel participation. The participation in the study does not entail any physical or psychological risks for me and my child. A cancellation of participation has no adverse consequences. I can contact the Heinrich-Heine-University in Düsseldorf (Valentin Wagner) at any time to ask questions.

(Place and Date) (Signature of parent)

Kernel density plots by Treatment

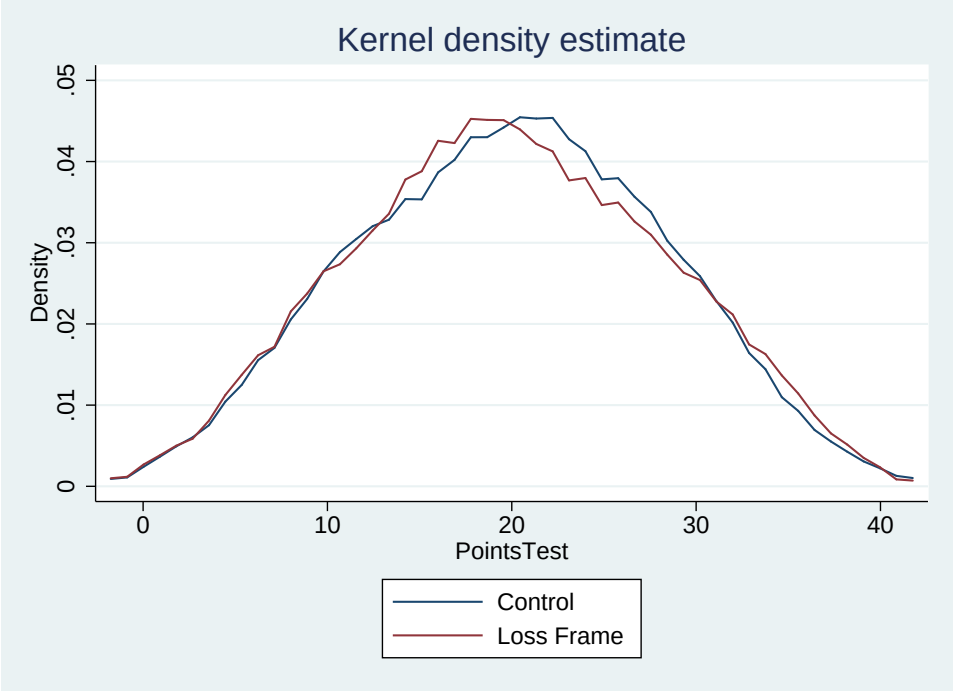
—not intended for publication—

Figure 10: Correct Answers: Loss Treatment vs. Negative Treatment



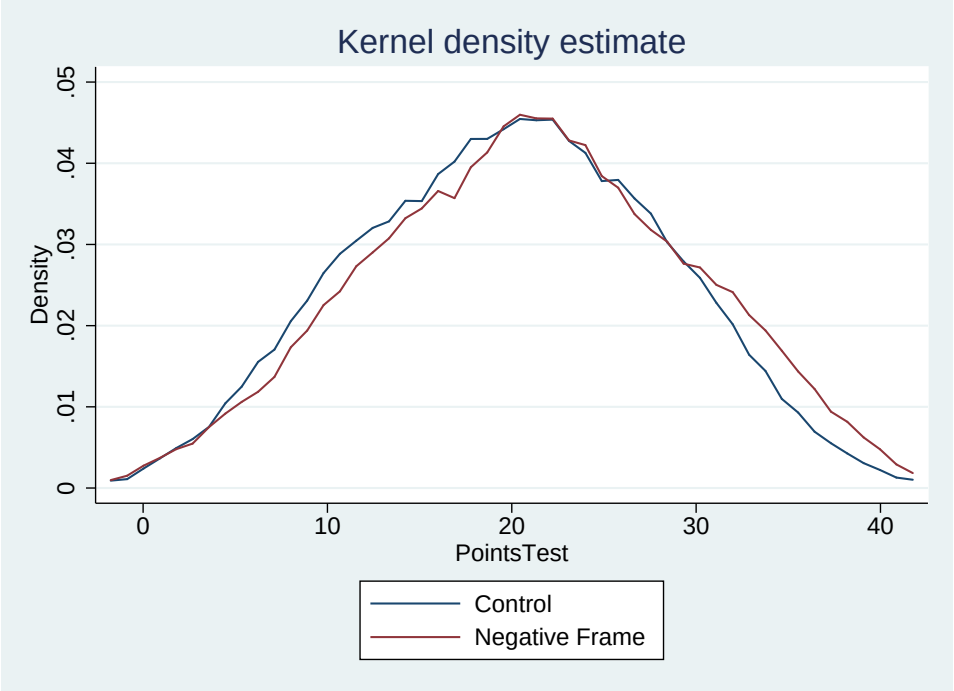
Note: This Figure presents Kernel density estimates for the number of correct answers for the Loss Treatment and the Negative Treatment.

Figure 11: Points: Control vs. Loss Treatment



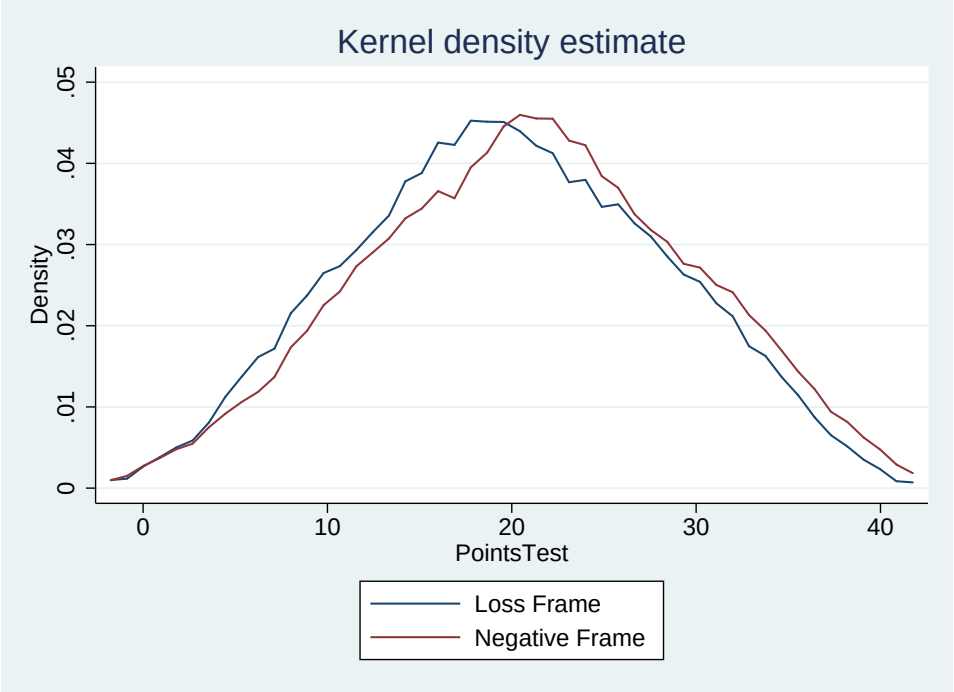
Note: This Figure presents Kernel density estimates for the number of points reached in the test for the Control Group and the Loss Treatment.

Figure 12: Points: Control vs. Negative Treatment



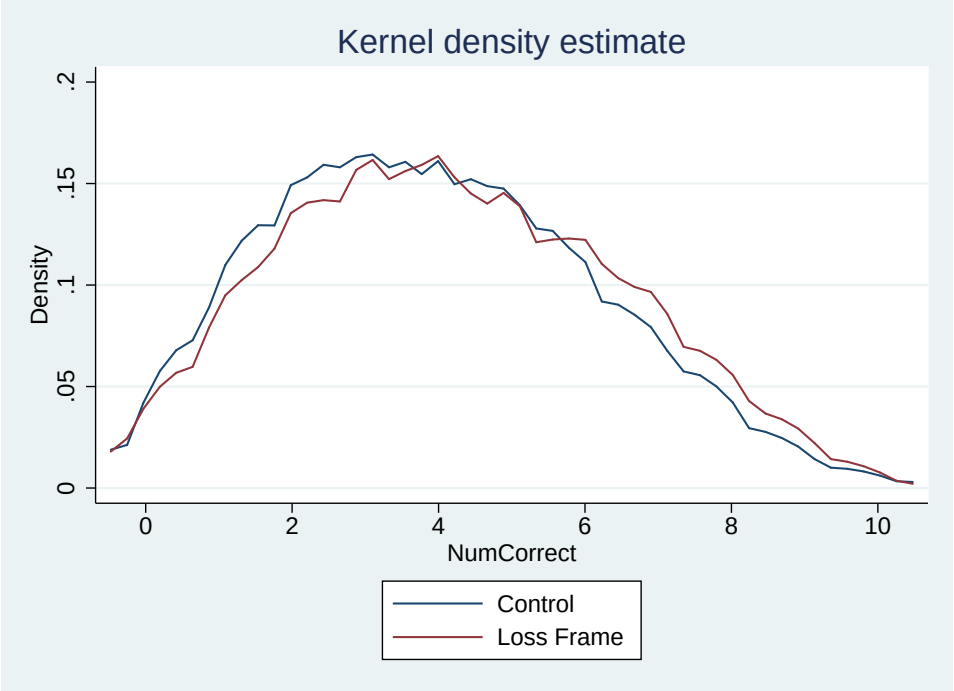
Note: This Figure presents Kernel density estimates for the number of points reached in the test for the Control Group and the Negative Treatment.

Figure 13: Points: Loss Treatment vs. Negative Treatment



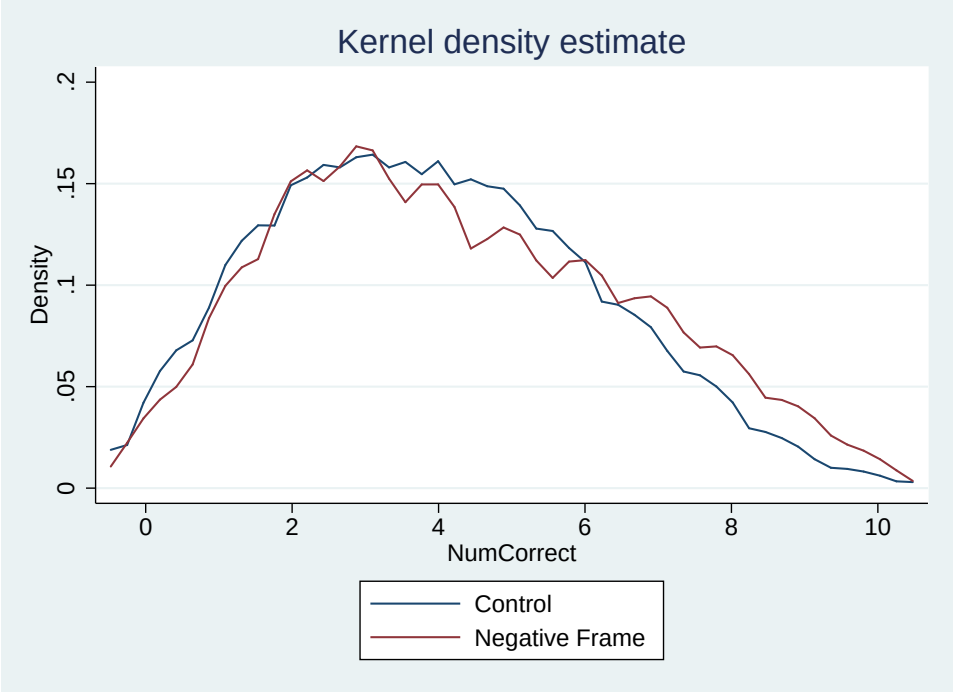
Note: This Figure presents Kernel density estimates for the number of points reached in the test for the Loss Treatment and the Negative Treatment.

Figure 14: Correct Answers: Control vs. Loss Treatment



Note: This Figure presents Kernel density estimates for the number of correct answers for the Control Group and the Loss Treatment.

Figure 15: Correct Answers: Control vs. Negative Treatment



Note: This Figure presents Kernel density estimates for the number of correct answers for the Control Group and the Negative Treatment.