Field Experiments in Behavioral Economics of Education

Inaugural-Dissertation

zur Erlangung des akademischen Grades eines Doktors
der Wirtschaftswissenschaften
(Dr. rer. pol.)
durch die
Wirtschaftswissenschaftliche Fakultät
der Heinrich-Heine-Universität Düsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

von:	Valentin Wagner, M.Sc. geboren am 12.12.1986 in Bonn
Erstgutachter:	Jun.-Prof. Gerhard Riener, PhD
Zweitgutachter:	Prof. Dr. Hans-Theo Normann
Abgabedatum:	14. Oktober 2016

Acknowledgment

I want to thank many people who supported me during the process of preparing this thesis. First, I want to thank my supervisor Gerhard Riener who set the cornerstone for this thesis on my very first working day in Düsseldorf during a joint rail travel to Bonn. His passion for experimental economics has always been inspiring and motivating. He has been an endless source of new ideas and creative solutions which shaped my own research and will impact all research that lies ahead of me. Second, I am very grateful to Hans-Theo Normann for always working with the door open and being an indispensable advisor on-site during Gerhard's time in Mannheim. Many thanks to both of you for your patience, constructive comments and for giving me the freedom to conduct research in a field which is not one of the core competencies of the institute.

This dissertation was written during my time at the Düsseldorf Institute for Competition Economics (DICE) at the Heinrich-Heine-University Düsseldorf and would have been impossible without Michael Coenen who offered me the position as a doctoral researcher. My research has benefited a lot from the many discussions with professors and colleagues at DICE as well as from presentations at various national and international conferences.

I want to thank all children, teachers, parents and principals who participated in my experiments. Furthermore, I thank the organizers of the Känguru-Wettbewerb, in particular Dr. Monika Noack, for granting me the permission to use test questions of the Känguru-Wettbewerb.

To conduct field experiments in schools without having knowledge about the teacher's perspective is a hopeless undertaking. I therefore thank Julia and Konrad for sharing their professional experience with me and for giving me valuable insights from the teacher's perspective. You contributed tremendously to the success of this dissertation by helping me to convince principals to participate in my studies, to get teachers understand my instructions and hence to successfully implement my research into the field. Moreover, I wish to thank Julia for having endured all frustrations and successes which have been associated with me writing this dissertation and for always having supported my urge to travel without ever complaining.

Finally, I want to thank my family. Even without having read countless studies in the economics of education literature about the importance of family background and the intergenerational transmission of education, I know that this dissertation would not have been possible without your help and support from the very beginning. I am particularly grateful to Ela and Bernhard who welcomed me every week again to stay at their place and gave me a second home on the other side (the "Schäl Sick") of the Rhine.

Valentin Wagner, October 2016

Contents

1 General Introduction 1

2 Peers or Parents? On the Signaling Value of Rewards in School 7

2.1 Introduction 8

2.2 Institutional Background and Selection of Test Incentives 13

2.2.1 Selection of Schools and Multiple-Choice Test 14

2.2.2 Survey and Selection of Test Incentives 15

2.3 Experimental Intervention 16

2.4 Data and Descriptive Statistics 20

2.5 Results 22

2.5.1 Peers or Parents? On the Selection of Incentives 22

2.5.2 Incentives and Test Performance 25

2.6 Further Explanations and Results 32

2.7 Conclusion 35

2.8 Appendix 37

3 Seeking Risk or Answering Smart? Framing in Elementary Schools 60

3.1 Introduction 61

3.2 Related Literature 63

3.3 Experimental Design 65

3.3.1 Selection of Schools and Choice of Testing Format 66

3.3.2 Treatments 66

3.3.3 Implementation 69

3.4 Hypothesis 71

3.5 Data and Descriptive Statistics 73

3.5.1 Attrition 74

3.6 Experimental Results 75

3.6.1 Framing and Test Performance 75

3.6.2 Who can be Framed? 81

3.7 Discussion 84

3.8 Conclusion 86

3.9 Appendix 88

4 Answering Strategies in Multiple-Choice Tests - Differences by School Types and Gender? 108

4.1 Introduction 109

4.2 Data Sources and Experimental Procedure 113

4.2.1 Experimental Design and Field Data . . . 114
4.3 Descriptive Statistics . . . 115
4.3.1 Experiment . . . 115
4.3.2 Field . . . 116
4.4 Results . . . 117
4.4.1 Hypothesis . . . 117
4.4.2 Differences in Skipping Test Items by Type of School? . . . 119
4.4.3 Gender Differences in Skipping Test Items and Test Performance 121
4.5 Discussion . . . 129
4.6 Conclusion . . . 134
4.7 Appendix . . . 135

5 Concluding Remarks 145

List of Tables

2.1 Comparison of Important Indicators: Experiment vs. North Rhine-Westphalia (in percent) 21
2.2 Chosen Incentives by Midterm Grade (in percent) 24
2.3 Multinomial Logit Model of Chosen Incentives 26
2.4 Treatment Effects 28
2.5 Treatment Effects by Gender 29
2.6 Chosen Incentives and Points in Test 30
2.7 Randomization Check 37
2.8 Treatment Randomization: Average Test Scores 38
2.9 Chosen Incentive by Gender and School Type (in percent) 39
2.10 Multinomial Logit Model of Chosen Incentives Cont. 40
2.11 Robustness Check—Treatment Effects without Covariates 41
2.12 Treatment Effects by Pupils' Midterm Grade 42
2.13 Treatment Effects by Pupils' Gender and Midterm Grade 43
2.14 Test Preparation by School Type 44
2.15 Ordinary Least Squares Regression of Covariates on Test Performance 45
2.16 Predetermined Incentives given in Survey 51

3.1 Treatment Overview 68
3.2 Comparison of Characteristics: Experiment vs. North Rhine-Westphalia (in percent) 74
3.3 Treatment Effects - Number of Correct Answers 77
3.4 Treatment Effects - All Outcome Variables 81
3.5 Treatment Effects by Ability 83
3.6 Sample Size by Gender, Grade and Treatment 88
3.7 Randomization Check 89
3.8 Attrition by Treatment 89
3.9 Treatment Effects - Number of Omitted Items 90
3.10 Treatment Effects - Share of Correct Answers 91
3.11 Treatment Effects - Total Points in Test 92
3.12 Treatment Effects without Control Variables- Correct, Omitted, Share and Points 93
3.13 Treatment Effects by Gender 94
3.14 Share of Correct and Answered Questions by Test Item 95
3.15 Robustness Check - Correct Answers, Omitted Answers, Points in Test 96
3.16 Treatment Effects by Midterm Grade 97

4.1 Descriptive Statistics - Field 117
4.2 Average Number of Omitted Questions, Probability of Success and Points achieved by Difficult Levels and by Incentivized 122
4.3 Number of Omitted Questions 124
4.4 Omitted Answers by Academic Year 127
4.5 Impact of Skipping on Test Performance 130
4.6 Descriptive Statistics on Skipped Answers 135
4.7 Raw Treatment Effects 136
4.8 Average Number of Omitted Questions, Probability of Success and Points achieved by Gender, Difficult Levels and by Incentivized . . . 137
4.9 Correct Answers by Academic Year 139
4.10 Omitted and Correct Answers by Prize Winner 140
4.11 Robustness Check - All Questions 141
4.12 Robustness Check - Easy Questions 141
4.13 Robustness Check - Middle Questions 142
4.14 Robustness Check - Difficult Questions 142

List of Figures

2.1 Time-line 18
2.2 Kernel Density Estimation Control vs. Incentivized 46
2.3 Survey Front and Back 49
2.4 Survey Answers 52
2.5 Example of No-Homework-Voucher 53
2.6 Example of Parent-Letter 54
2.7 Picture of Medal 55
2.8 Instructions for Teachers - First Mailing [Translated from German] . 56
2.9 Instructions for Teachers - Second Mailing [Translated from German] 57
2.10 Teacher Questionnaire [Translated from German] 58
2.11 Pupil Questionnaire [Translated from German] 59

3.1 Average Number of Omitted Answers and Share of Correct Answers . 79
3.2 Average Number of Omitted Answers and Share of Correct Answers by Gender 98
3.3 Kernel Density Plot: Points after Five Questions (Q1-Q5) 98
3.4 Kernel Density Plot: Final Points of Pupils who accumulated 20 Points at Q5 99
3.5 Teacher Instructions—First Letter [Translated from German] 100
3.6 Teacher Instructions—Second Letter [Translated from German] 101
3.7 Teacher Questionnaire [Translated from German] 102
3.8 Student Questionnaire [Translated from German] 103
3.9 Consent Form to be Signed by Parents (Translated from German) . . 104
3.10 Correct Answers: Loss Treatment vs. Negative Treatment 105
3.11 Points: Control vs. Loss Treatment 105
3.12 Points: Control vs. Negative Treatment 106
3.13 Points: Loss Treatment vs. Negative Treatment 106
3.14 Correct Answers: Control vs. Loss Treatment 107
3.15 Correct Answers: Control vs. Negative Treatment 107

4.1 Number of Answered Questions and Points Received in Test 120
4.2 Number of Answered Questions and Points Received in Test by Gender128
4.3 Distribution of Omitted Questions 138
4.4 Kernel Density Estimation Incentive vs. No Incentive 143
4.5 Kernel Density Estimation Male vs. Female 144

Chapter 1

General Introduction

Education and the accumulation of human capital is closely linked to the economic growth of modern knowledge-based economies and individuals' lifetime earnings. Hanushek and Wößmann (2012) study the role of cognitive skills on economic growth and their results suggest that one-standard-deviation higher cognitive skills of a country' s workforce is associated with approximately 2% higher annual growth in per capita GDP. Oreopoulos (2007) analyzes how compulsory schooling affects subsequent outcomes and shows that individuals' wealth increases by about 15% with an additional year. Similar effect sizes are also found in other studies. Hanushek et al. (2015) find that a one-standard-deviation increase in numeracy skills is associated with an average increase in hourly wages of around 18%, Harmon and Walker (1995) estimate that the return to schooling in the UK is of the order of 16% and more recently Bhuller et al. (2014) find that the internal rate of return of schooling in Norway is of around 10%.[1] Moreover, increases in the return to post-secondary education account for most of the past growth in wage inequality (Lemieux, 2006).

Higher levels of education are associated with higher wages in the first place. However, education also generates positive externalities on various other areas of life and researchers have in particular used changes in compulsory schooling laws to investigate the causal links on outcomes such as health, crime, teenage pregnancy and political behavior. On health, there seems to be a strong connection between higher levels of education and better health. This is in part due to higher incomes and occupational choice, but Cutler and Lleras-Muney (2006) suggest that increasing levels of education also lead to different thinking and decision-making patterns which in turn affect individuals' health status and health behavior. Moreover, the effect of education on health seems to be at least as great as the effect on income (Feinstein et al., 2006). In terms of crime, the seminal paper by Lochner and Moretti (2004) shows that participation in criminal activities can be reduced by increasing the years of compulsory schooling which also results in substantial social savings.[2] Another area of life that is affected by education is fertility, i.e. teenage pregnancy. Monstad et al. (2008) find that education influences the intertemporal choice of having children in Norway. While an increase in education leads to postponement of first births away from teenage motherhood towards later years, the overall decision of having children is not affected. There is no evidence that more education results in more women remaining childless or having fewer children. Black et al. (2008) find a similar relationship between compulsory schooling and fertility for Norway and furthermore also for a different institutional environment like the US. Regarding political behavior, education seems to create benefits to society. Milligan et al. (2004) find that better educated adults are more likely to follow election campaigns in the media, discuss politics with others, associate with a political group, and work on community issues. Moreover, there is also evidence that schooling could affect multiple dimensions of skills. Oreopoulos and Salvanes (2011) argue that schooling may shape individuals' non-cognitive skills such as patience, long-term thinking, im-

[1] Ashenfelter et al. (1999); Oreopoulos (2006); Angrist and Krueger (1991); Card (1999) and Trostel et al. (2002) use different estimation methodologies and also show a positive rate of return on schooling.

[2] See also Machin et al. (2011) on the positive effects of increased compulsory schooling on (property) crime reductions.

proves trust, increases goal-orientation and reduces the likelihood to engage in risky behavior.[3] Nevertheless, all studies examining the impact of education on non-pecuniary effects have to handle the challenge to exclude the effects of the higher incomes brought about by schooling and to take into account that a higher amount of schooling may be correlated with family background and socio-economic status (Oreopoulos and Salvanes, 2011).

It seems puzzling that some individuals invest poorly in their own education despite the many positive effects of education on almost all aspects of life. This underinvestment could be explained by individuals (i) overly discounting the future, (ii) having time-inconsistent preferences, (iii) underestimate the (financial) return to education (Gneezy et al., 2011) or by (iv) individuals' unawareness of their own production function (Cunha and Heckman, 2007). Furthermore, there might be a lack of information. Jensen (2010) argues that the *perceived* returns to education are important for the schooling decision and that families, especially in low-income countries, are not well informed about these returns. The author shows that providing information on the return to education increases the perceived earnings and thus, treated families completed more years of schooling.

To succeed in the educational system is important and researchers have therefore focused to a large extend on the ex-post evaluation of school reforms and changes in the institutional setting. This comprises inter alia class-size reduction (Hoxby, 2000), co-teaching (Andersen et al., 2015), mixed-age classes (Veenman, 1995; Lindström and Lindahl, 2011), tracking (Betts, 2011) and shortening secondary school duration— "G8-Reform" (Büttner and Thomsen, 2015). However, another and potentially the most important input for success in education is pupils' *motivation*.[4] Motivation is linked to pupils' positive or negative attitude towards schooling and to motivate pupils to invest in their education is part of teachers' daily work.

I am therefore interested to study pupils in their natural learning environment, to what extend insights from behavioral economics can be transferred to the educational sector, how pupils can be motivated to invest in their own education and to provide teachers with potential cost-effective *"tools of motivation"*. To do so, I conducted field experiments in elementary and secondary schools in North-Rhine Westphalia, Germany. Researchers can opt for different empirical methodologies and a common approach to isolate the causal effect of schooling on the outcome variable of interest is to exploit data from "natural" experiments or to use identification strategies such as regression-discontinuity designs, (propensity score) matching methods, differences-in-differences estimation or instrumental variables (Dolan and Galizzi, 2014).[5] Nevertheless, the methodology of field experiments is often considered as the "gold standard" for program evaluation in education research (Sadoff, 2014) and the greater use of field experiments could lead to the more efficient use

[3] Cognitive skills are usually identified with intelligence and the ability to solve abstract problems whereas non-cognitive skills are personality traits that are weakly correlated with measures of intelligence (Brunello and Schlotter, 2011).

[4] See Ryan and Deci (2000) on extrinsic and intrinsic motivation.

[5] See also Schlotter et al. (2011) for a discussion of econometric methods for causal evaluation of education policies.

of scarce resources (Dolan and Galizzi, 2014).[6] The advantage of the experimental methodology is that identification assumptions are less severe and the core of each experiment is appropriate *randomization*. Random assignment to the intervention can be treated as an instrumental variable that is exogenous by definition (List, 2007) and ensures that there are no underlying differences on average between the treatment and control groups. Therefore, random assignment solves the selection bias (Duflo et al., 2007)[7] and any differences in outcomes between the treatment and control groups can be attributed to the intervention itself, allowing to *causally* identify the impact of a given program (Sadoff, 2014). Furthermore, field experiments often represent a mixture of control and realism which is usually not achieved in the lab or with uncontrolled data (Levitt and List, 2009). However, experimentalists have to deal with other obligations such as the optimal number and arrangement of subjects into experimental cells[8] and a limitation of field experiments — unlike lab experiments — might be replicability (Falk and Heckman (2009) discuss the advantages and limitation of lab experiments by comparing them to field experiments and non-experimental data).

In this thesis, I conduct field experimental in schools and focus on two possibilities out of many which could change pupils' motivation in school: (i) non-monetary incentives and (ii) framing manipulations. However, motivation could also be enhanced by setting intermediate deadlines (this is ongoing research with Gerhard Riener, Heinrich-Heine-Universität Düsseldorf) to overcome the problem of procrastination, or teachers could give pupils feedback about their past performance. The latter is ongoing research with Mira Fischer (Universität zu Köln). In a field experiment in secondary schools, we provide pupils with feedback about their past performance either immediately or three days before a high-stakes exam — the last math exam of the semester. We vary whether pupils receive static feedback — the rank in the last exam — or a dynamic feedback — the change in the rank between the first and second exam. Furthermore, we investigate whether these different types of feedback affect motivational beliefs such as locus of control, self-esteem or math self-efficacy (the beliefs regarding the own power to affect mathematics "situations"). The preliminary analysis of our data suggests that pupils receiving a negative feedback (static as well as dynamic) a few days before the exam improve their performance compared to pupils with positive feedback and compared to pupils without any form of feedback.

I analyze how motivation and academic performance can be manipulated in the educational sector in the three following chapters:

Extrinsic financial incentives are a "natural" resource of economists to solve problems in motivation. However, this type of incentives can be very costly, not feasible for policy makers and teachers along with parents are mostly critical about

[6] According to Harrison and List (2004), field Experiments can be categorized broadly into three main types: artefactual, framed, and natural field experiments.

[7] The selection bias can occur if individuals self-select into and hence are not randomly assigned to treatment groups which could result into pre-existing differences between the control and treatment groups.

[8] See List et al. (2011) and Duflo et al. (2007) on randomization techniques.

"cash for grades". **Chapter 2** entitled **"Peers or Parents? On the Signaling Value of Rewards in School"** (co-authored with Gerhard Riener) therefore aims at identifying to whom pupils want to signal their academic achievement to be able to better tailor and to increase the effectiveness of *non-monetary* incentives. To do so, we conducted a field experiment in high- and low-achieving schools in Germany with more than 2.000 pupils. Incentives were provided for self-improvement in a mathematical test and were either predetermined or self-selected. In the latter (the Choice Treatment), pupils could choose one out of four incentives. These non-monetary incentives were selected based on a survey conducted prior to the experiment and differed with respect to the principal target audience — peers or parents. In particular, pupils could choose between a medal, a parent-letter, a homework voucher or a surprise. To test the effectiveness of rewards on pupils' performance, academic achievements in the Choice Treatment are then compared to two predetermined treatment conditions (the Fixed-Medal and Fixed-Letter Treatments) and to pupils who were not eligible to receive a reward (the Control Group). We find that pupils with lower maths grades choose a reward with a higher signaling value to their parents while high-achieving pupils tend to choose to signal their academic achievements to their peers. We find no differences in the signaling decision by gender or school type (Vocational vs. High Schools). However, the effectiveness of the *predetermined* incentives on test performance differs by school type. Test performance decreases significantly for pupils in High Schools but not for pupils in Vocational Schools. In contrast, when allowing for *choice* over the incentive, we do not observe a decrease in pupils' performance in High Schools and moreover, an increase in pupils' willingness to prepare for the test.

Insights from behavioral economics and its value for applications in policy-making have been increasingly recognized by governments in recent years. In 2010, the European Commission set up the "Framework Contract for the Provision of Behavioral Studies (FCPBS)[9]", in 2014 the US government assembled the "Social and Behavioral Sciences Team[10]", the World Bank officially launched its "Global Insights Initiative (GINI)[11]" in 2015 and a number of European countries (i.e. UK, Netherlands, Germany, France and Denmark) installed specialized behavioral insights teams.[12] Despite the increasing application in policy-making, behavioral concepts have been rarely applied to the educational sector although they constitute a promising source for motivating pupils. **Chapter 3** entitled **"Seeking Risk or Answering Smart? Framing in Elementary Schools"** therefore tests the motivational power of framing effects on pupils' decision making in a multiple-choice test, in particular, how loss and gain framing affects the quantity and quality of decision. In a field experiment in elementary schools, 1.377 pupils were randomly assigned to one of three experimental conditions: (i) gain frame (Control Group), (ii) loss frame (Loss Treatment) and (iii) gain frame with a downward shift of the

[9] `http://is.jrc.ec.europa.eu/pages/BE/BEindex.html`

[10] `https://sbst.gov/`

[11] `http://www.worldbank.org/en/programs/gini`

[12] The European Commission gives a recent overview about European countries applying behavioral insights to policy (`http://publications.jrc.ec.europa.eu/repository/bitstream/JRC100146/kjna27726enn_new.pdf`).

point scale (Negative Treatment). According to prospect theory (Kahneman and Tversky, 1979), individuals evaluate a loss approximately twice as much as an equal gain if they are loss averse and therefore pupils should increase their performance if they are endowed with the maximum score. On average, I find that pupils in both treatment groups answer significantly more questions correctly compared to the "traditional grading". This increase is driven by two different mechanisms. While pupils in the Loss Treatment increase significantly the quantity of answered questions — seek more risk — pupils in the Negative Treatment seem to increase the quality of answers — answer more accurately. Moreover, differentiating pupils by their initial ability shows that a downward shift of the point scale is superior to loss framing. High-performers increase performance in both treatment groups but motivation is significantly crowded out for low-performers in the Loss Treatment.

Performance in a test depends on pupils' motivation but could also be influenced by the testing format. Testing formats can be broadly divided into open-ended questions and multiple-choice questions. Multiple-choice university entrance exams determine access to higher education in many countries, which is one important prerequisite for later employment possibilities. The application of this testing format is problematic if it favors answering strategies of certain groups in the population. Recent experiments have identified guessing as one reason for gender differences in performance (Pekkarinen, 2015; Baldiga, 2014), but as promotion within the educational system should depend on actual knowledge and not on how knowledge is assessed, this poses a challenge for general multiple-choice tests. **Chapter 4** entitled **"Answering Strategies in Multiple-Choice Tests - Differences by School Types and Gender?"** (co-authored with Gerhard Riener) investigates whether answering strategies in multiple-choice tests differ between school types, gender and school grades. We address three questions: First, whether pupils in different school types apply different answering strategies? Second, whether the gender gap in guessing — boys are typically found to guess more often than girls — exists across social strata? Third, whether the gender gap exists over all school grades? To answer the first two questions, we exploit data from the randomized field experiment in *Chapter 2*. Pupils in secondary schools in Germany can be differentiated by social background and intellectual ability as measured by the school type (Vocational and High Schools). Our experimental data are complemented by using aggregate data of a nationwide test with more than 780.000 participants of grades 3 to 12 to shed light on the third question. We find that pupils in High School skip more answers than their counterparts in Vocational School but that they obtain higher test scores by answering more accurately. Results on gender differences reveal a gender gap in skipping math tasks only for pupils from higher socio-economic families and only if questions are difficult. However, this gap can be closed by providing extrinsic rewards for performance suggesting that the gender gap in skipping test items could be in line with a stereotype-threat explanation. Moreover, gender differences are found in all school grades and tend to increase over the years.

Chapter 2

Peers or Parents? On the Signaling Value of Rewards in School

Co-authored with Gerhard Riener

Contributions of Valentin Wagner

- Development of research idea and literature review
- Establishing contact to schools and preparatory talks
- Design of experiment and expiration
- Data preparation, descriptive analysis and graphs
- Treatment effect estimation and robustness checks
- Textual contributions in all sections of the paper

____

Gerhard Riener

2.1 Introduction

Pupils often lack the motivation to study mathematics, although mathematical skills yield a large economic premium and are an important prerequisite for later employment possibilities and wages (Hanushek et al., 2015).[1] Pupils might under-invest in their own mathematical education because they are not aware of their own production function (Cunha and Heckman 2007), they may underestimate the return on education (Oreopoulos 2007; Gneezy et al. 2011)[2] or are afraid of not being accepted by their peers by performing in a manner that is not consistent with the group's expectations (Akerlof and Kranton 2005; Austen-Smith and Fryer 2005). Even if pupils recognize the individual importance of mathematical eduction, it has positive externalities, which may lead to sub-optimal investment. It is thus crucial for educational policy to understand how pupils are motivated to enhance their performance and improve their attitude toward mathematics.

An economist's natural recourse to increase performance is through financial incentives. However, implementing monetary incentives in an educational context entails at least three obstacles: (i) it is potentially more cost-intensive than the status quo,[3] (ii) there is low acceptance for "cash for grades" by teachers and parents who think that education has value and entails motivation in itself and (iii) it raises ethical issues. Moreover, research on financial incentives in schools has revealed mixed results (Fryer 2011; Bettinger 2012; Levitt et al. 2016) as these incentives may crowd out internal curiosity and motivation to acquire new knowledge, which underlines the second concern.

In addition to financial incentives, non-monetary rewards that use public recognition of success may be effective. Public recognition has been identified to influence motivation and performance (Bursztyn and Jensen 2015) but the role of the target audience—in particular in educational settings—remains unclear. Recent work in labor and personnel economics provides evidence that recognitional incentives have

[1] Hanushek et al. (2015) show in a study across 22 countries that a one standard deviation increase in numeracy skills is associated with an average increase in hourly wages of 17.8% (see also Niederle and Vesterlund [2010]; Goodman [2012]).

[2] Cunha and Heckman (2007) develop a model of skill formation with multiple stages of childhood in which inputs at different stages are complements and self-productivity of investment is present. Due to dynamic complementarity—which means that the marginal productivity of investment depends on the level of skills produced by previous investments—it may be difficult for individuals to know their educational production function. Oreopoulos (2007) evaluates the impact of compulsory schooling on dropout rates. He finds that lifetime wealth increases by about 15% with an extra year of compulsory schooling. According to Oreopoulos (2007), dropouts likely forgo substantial gains to lifetime wealth because adolescents ignore or heavily discount future consequences when deciding to drop out of school. In the 1990 Eurobarometer Youth Survey, more than 50% of 16 to 25-year olds leaving school at the minimum age indicated that their reason for dropping out was lack of interest or that they saw no point in going on.

[3] For example Fryer (2011) distributed a total of \$9.4 million (approx. \$348.15 per pupil—treated and untreated) and \$650.000 (roughly \$385 per treated student) were awarded by Angrist and Lavy (2009). Fryer (2011) tests the effectiveness of financial incentives in Dallas, New York and Chicago. He finds that the incentives offered for educational outputs (such as better grades) are less effective than incentives for educational inputs, such as attendance or reading books. Angrist and Lavy (2009) offered cash awards to students in Israel who passed their exams as part of an attempt to increase certification rates among low-achievers. These cash awards led to an increase in certification rates for girls but not for boys because girls devoted extra time to exam preparation.

the power to keep up or increase workers' motivation (Kosfeld and Neckermann 2011; Kube et al. 2012) and might also work in schools because children are often higher motivated by short-run rewards than less tangible long-run rewards (Chelonis et al. 2004; Bettinger and Slonim 2007).[4] Moreover, delegating the choice of the incentive to the recipient has been shown to have additional positive effects in experimental labor markets (Charness et al. 2012). Thus, incentives that aim at recognizing the achievement of a student within the classroom or that inform a student's parents may provide a simple and cost effective way to circumvent the problems of financial incentives. These types of incentives are accepted and frequently used by teachers (Caffyn 1989)—who are, of course, important stakeholders in the implementation of the policy—and are politically feasible. Furthermore, pupils' empowerment by letting them participate in the learning environment is a positively valued feature.

Thus far, there is little knowledge to which audience—peers or parents—pupils want to reveal their educational achievement and little research has focused on the effectiveness of public recognition of academic merit that may be a viable alternative in the educational sector. However, extrinsic non-monetary incentives do not come entirely free either as there is the danger of hidden costs. First—as with monetary incentives—there is the potential of crowding out intrinsic motivation if rewards are too low-powered or not properly designed (Gneezy and Rustichini 2000). Second, peer group effects may gain importance as the performance or changes in performance will be made public (Bursztyn and Jensen 2015); thus, depending on the audience, recognition can have ambiguous effects (see also Austen-Smith and Fryer [2005], on "Acting White").

We want to observe pupils' choice of recognitional incentives that differ with respect to the target audience—a medal awarded in front of a student's peers and a letter sent to a student's parents. The delegation of choice over incentives to pupils represents a major novel contribution of our paper. Moreover, we test how these recognitional incentives work on pupils' test performance for different school types (Vocational vs. High Schools). These kind of incentives might potentially retain some of the power and simultaneously mitigate some of the problems of cash incentives. Therefore, we provide pupils in secondary schools with (non-monetary) public recognition incentives for individual improvement in a mathematical test. Rewards are exogenously determined in two of the treatments (Medal and Letter Treatment), whereas the choice of the incentive is delegated to the pupils in the Choice Treatment. We are also interested in how public recognition incentives and delegation interact with gender and socio-economic background—as measured by the school type. Since reputational effects are grounded in social customs within the classroom and the family environment—and because we do not expect that the effect of recognitional incentives will change in the short or medium run—we focus on short-run effects.[5]

[4] A recent literature review by Koch et al. (2015) offers an overview of other approaches in behavioral economics—such as self-control, willingness to compete, self-confidence and the influence of the environment—which can explain educational investment decisions and outcomes in education.

[5] Altering the underlying attitudes towards educational achievements would require different types of (and potentially more costly) interventions; we therefore test the immediate applicability of an incentive scheme based on recognition.

The experiment is conducted within the German school system which is characterized by early school tracking and these tracking choices are highly correlated with pupils' socio-economic background (Dustmann 2004; Ditton 2007; Paulus and Blossfeld 2007).[6] We therefore distinguish between High Schools and Vocational Schools; pupils attending High School belong, on average, to families of higher socio-economic status. The final examination of High Schools entitles students to apply to University whereas Vocational Schools usually prepare pupils for vocational jobs. Our sample consists of younger pupils in all school types (fifth and sixth graders; 11-years old, on average) because non-monetary incentives tend to work better than financial rewards for this age group. As has been argued by Levitt et al. (2016), younger children who are less familiar with cash may be more responsive to non-financial rewards than older students who are more familiar with cash. Moreover, increasing educational inputs in younger ages is promising, as these inputs are likely to complement skill formation in later stages of education (Cunha and Heckman 2007). Our sample of selected schools matches important indicators of school success on the county level of North-Rhine Westphalia. However, although difference in socio-economic status is one key difference between pupils of Vocational Schools and High Schools, they are likely to also differ on other measures. Therefore, differences between school types resemble suggestive evidence on differences in socio-economic background but can not be claimed to be the only one.

We present three sets of results. *First*, the overall result in which the signaling decision—choice of the target audience—depends on pupils' ability. We find that low-performing pupils in the Choice Treatment are significantly more likely than high-performing pupils to choose the parents letter and that high-performers tend to choose more often the medal compared to low-performers. *Second*, the effectiveness of public recognition incentives is shown to depend on school type. Pooling over gender and age, we find that the effect on performance among pupils from higher socio-economic families (High School pupils) is negative for *predetermined* rewards (the Medal and Letter Treatment), while we observe no significant effect on performance for pupils from families with lower socio-economic background (Vocational School pupils).[7] *Third*, we find no decrease in educational achievement if pupils in High Schools are free to choose their incentive.[8] We also find that the delegation over the incentive scheme significantly increases pupils' (self-reported) willingness to prepare for the test.

Although there is ample but mixed evidence on the effectiveness of incentivizing teachers using pupils' performance (Lavy 2002; Springer et al. 2011; Fryer 2013; Muralidharan and Sundararaman 2011)[9], few studies systematically evaluate the role

[6] School types in Germany differ in their education of teachers because universities offer two different degree programs with different focus areas. Furthermore, students becoming High School teachers typically have a higher High School graduation score.

[7] The difference between the school types is significant for the Letter Treatment.

[8] We can formally reject that the insignificant positive estimate in the Choice Treatment is not equal to the significant negative effects of the Medal and Letter Treatment.

[9] While teacher incentives in developing countries have shown promising results (Lavy 2002; Muralidharan and Sundararaman 2011), experiments in the US suggest that teacher incentives are ineffective (Springer et al. 2011; Fryer 2013). For an overview of the effectiveness of performance-based pay systems on teachers see Neal (2011).

of pupils' target audience and the effects of incentivizing pupils with public regognition incentives. Furthermore, the different *level of achievement* regularly achieved at different schools has largely been neglected because previous experimental studies have mainly focused on deprived schools.[10]

On non-monetary incentives Levitt et al. (2016) compare an in-class trophy awarded for good performance on a test to monetary rewards. The authors show that this non-monetary incentive has larger effects than financial incentives for younger pupils. Furthermore, these authors find that incentives work better when the bonus is paid immediately instead of delayed by a month.[11] Jalava et al. (2015) analyze the effect of grading methods (rank-based grading vs. criterion-based grading)[12] in Swedish schools and provide non-monetary incentives—a certificate and a prize (refillable pencil). Their findings are comparable to ours as pupils are also in grade 6 and typically twelve years old. Pupils in the "Certificate-Treatment" were promised a certificate if they exceeded the criterion-based score for A-B (18 points or more). An additional treatment rewarded pupils if they were among the top three performing pupils in their class. Jalava et al. (2015) find that the effectiveness of non-monetary incentives differs across the test score distribution and with respect to gender. Boys and girls increase their performance equally in the rank-based grading treatment, but girls also respond strongly to the certificate reward. The non-monetary incentives primarily work positively for pupils in the middle quartiles of the ability distribution and crowd out intrinsic motivation for low-ability pupils.

Closely related to the literature on non-monetary rewards is the literature on effort response to private versus public rewards as the way rewards are distributed may change their motivational power (Lacetera and Macis 2010; Neckermann and Frey 2013). Furthermore, Ariely et al. (2009) show that monetary incentives depend on visibility; for prosocial activities, monetary incentives are more effective if these activities are private rather than observed publicly.

[10] A notable exception is Angrist and Lavy (2009). Cash awards were provided for low-achieving high school pupils in Israel. However, the sample consisted of 40 nonvocational high schools with the lowest Bargut (matriculation certificate) ratings in a national ranking. For further studies on monetary incentives in the educational system see Fryer (2011), McEwan (2015), Blimpo (2014), Bettinger (2012).

[11] In a university setting, Chevalier et al. (2014) conducted a controlled field study among first-year undergraduate economics students that varied the incentives rewarding effort on a quiz. Incentives included, inter alia, additional educational material, a book voucher for the top performer or the quiz grade counted for 2.5% or 5% towards the final grade of the course. Chevalier et al. (2014) find that assessment weighting is highly effective in improving quiz participation, which improves performance on the final exam. Each additional quiz improved grades by 0.15 of one standard deviation. Bigoni et al. (2015) tested the effects of non-monetary incentives (extra points on the next exam) on university students (depending on their performance in previous tests, students could earn bonus points for the final exam). They employ both a cooperative and a competitive treatment. In the cooperative treatment, a student's test score was increased by one extra point if her partner's score was sufficiently good. In the competitive treatment, a student's mark in a test was increased by two extra points if her score resulted was higher than her partner's. Although women did not respond to the incentive at all, men—and particularly low-ability men— performed better in the competitive treatment. Bigoni et al. (2015) find no difference between the control and cooperative treatment.

[12] In the criterion-based grading treatment, pupils received grades on an A-F scale based on their performance, whereas in the rank-based grading treatment, the top three performing pupils within a class received a grade of A.

Thus far, there are only a few experiments that evaluate the role and characteristics of the target audience of public rewards. Moreover, there is no field experiment, to our knowledge, that analyzes the effect of an endogenous choice over rewards–and hence the target audience—in schools, although it appears that incentives such as recognition are budget-neutral and are appreciated within the pedagogical community. Bursztyn and Jensen (2015) is the closest to our study; the authors analyze peer effects in one natural and one field experiment when performance or investment in education is either observable or kept private. In the natural experiment, top students' performance declined by about 40 percentage points when revealed to the class, whereas lower performing students improved slightly.[13] In their field experiment, Bursztyn and Jensen (2015) find that investment in education depends on to whom the investment decision would be revealed. Students were offered complimentary access to an online SAT preparatory course and students' decision was either kept private or revealed to classmates. Students in honors classes were more likely to sign up for the preparatory course when the decision was made public rather than kept private, while students in non-honors classes were less likely to sign up if the decision was made public.[14] Bursztyn and Jensen (2015) vary the peer group composition but neglect the role of parents as big stakeholders of recognition although there is evidence that incentives with signaling value targeted towards parents is promising to increase pupils' behavior and performance. Avvisati et al. (2014) show that motivating parents to become involved in their children's education can change pupils' behavior. Particularly with parental involvement, pupils developed more positive behavior and attitudes in school, notably in terms of truancy and disciplinary sanctions (see also Fryer et al. [2015] and the literature therein on the importance of parental inputs).

To summarize, publicly awarded rewards have the power to increase workers' motivation but in an educational setting the motivational effect seems to depend on the target audience (peer group). Although programs involving parents in school activities show favorable results to change pupils' behavior, the effect of rewards with signaling value to parents on pupils' achievements is unclear so far. Furthermore, there is little evidence on pupils' preferred target audience and the value of delegation in education.

Our contribution to the literature is fourfold. First, we contribute to the literature on the signaling value of rewards. Giving pupils the opportunity to choose an incentive, we can analyze to whom—peers or parents—pupils want to signal their achievement in education. This signaling value is likely to change if the peer group composition changes (Bursztyn and Jensen 2015). Providing public recognition incentives to pupils from different school types, we can examine the correlation between signaling value and socio-economic status.

Second, we contribute to the growing literature on the effects of empowerment on human performance by giving pupils the flexibility and freedom to choose their

[13] On average, performance declined by 24%. The names of the top three scorers in the class were displayed on leaderboards.

[14] Students taking both honors and non-honors classes were 15 percentage points less likely in non-honors classes to sign up if the decision was public rather than private but were 8 percentage points more likely in honors classes to sign up if the decision was public.

reward beforehand. Individuals who choose their wage payment exhibit higher performance in experimental labor markets (Charness et al. 2012) and individuals who choose an activity are likely to perform and cooperate better than those who are assigned to an activity (Bo et al. 2010; Sutter et al. 2010). Thus far, little is known of the effect on effort when people are free to determine their compensation scheme (see Mellizo et al. [2014] for a recent study). To our knowledge, no study has yet applied this method to the educational sector.

Third, we examine the effect of public recognition incentives on academic achievements in high- and low-performing school types, which are mainly but not exclusively differentiated by the socio-economic status of their student bodies. Until now, the literature has focused primarily on incentives provided in deprived schools. However, it is also of interest to policy makers, educators and parents to learn how pupils of high-achieving schools react to these incentives.

Fourth, we extend the literature on non-monetary incentives in school by extending and comparing the set of non-monetary incentives. Thus far, a trophy (Levitt et al. 2016), a certificate and a refillable pencil (Jalava et al. 2015) have been tested. This study is complementary to (Jalava et al. 2015), as we use an incentive scheme that attempts to avoid crowding out of motivation for low-performing pupils in that pupils compete against their past scores and are awarded for self-improvements.

The paper is organized as follows. In Section 2.2 we give background information on the German school system and on the selection of test incentives. Section 2.3 explains the experimental design and Section 2.4 presents the data. In Section 2.5, we present our results, which are discussed in Section 2.6. Section 2.7 concludes.

2.2 Institutional Background and Selection of Test Incentives

The German school system offers a good setting in which to analyze pupils' signaling decision and the impact of public recognition incentives on performance in different institutional environments because it children are segregated into high- and low-performing groups at the age of ten. We run our experiment on pupils in grades 5 and 6 as these grade levels serve to test, promote and monitor pupils and to decide in cooperation with parents on the suitability of pupils for the chosen type of school: suitability is assigned with successful promotion after grade 6. We provide a more detailed description of the German school system in Appendix 2.8.

Peer composition in the classroom is determined by a tracking system, which begins after grade 4 of elementary school. It is therefore important to understand the transition process from elementary school to secondary education to recognize how it translates into the social composition of pupils between school types. Furthermore, secondary school track choice has major effects on subsequent educational achievements and labor market outcomes (Dustmann 2004; Dustmann et al. 2016).

Parental social status has a significant twofold influence on the choice of school type (Gresch et al. 2010). First, the social status of parents directly influences school performance in elementary school and hence the transition recommendation. Pupils from families with higher socio-economic status are more likely to be recommended to High School based on their better school performance. Second, parents from a privileged background put more emphasis in sending their children to academically advanced school types than parents with low socio-economic status (see Ditton [2007]; Paulus and Blossfeld [2007]). These parents are also more likely not to follow the school recommendation and to enroll their child at a school type of their original choice if they do not receive the desired transition recommendation. For example, Dustmann (2004) shows that parental background is strongly related to children's secondary track choice. Furthermore, Jonkmann et al. (2010) provide a more recent and detailed overview about the dependency between parents' educational background and children's tracking decision in Germany. They show that approximately 62% of pupils whose parents have the highest school graduation also attend High School. In comparison, approximately 35% of pupils whose parents have middle-level school graduation and only 14% of pupils whose parents have the lowest school graduation attend High School.

2.2.1 Selection of Schools and Multiple-Choice Test

Schools Using a list of schools that is publicly available from the Ministry of Education of North Rhine-Westphalia (NRW), we contacted 170 schools in the cities of Bonn, Cologne and Düsseldorf, which represent 9.5% of secondary schools in NRW.[15] Contact was first established via email and posted letter on November 19, 2013. As the average information transfer in school takes about two weeks (according to informal inquiries within schools), we contacted the schools again on December 9, 2013. About 33% of all schools responded, and 28 schools replied positively and agreed to a preparatory talk.[16] In these preparatory meetings, the experimental design was explained to at least one teacher per school and lasted about 30 minutes. Finally, 25 schools totaling 89 classes agreed to participate in the experiment.

Multiple-Choice Test We received permission to use questions from a mathematics competition test (*Känguru-Wettbewerb*) that is administered throughout Germany and in over 50 other countries. The mathematical test consisted of 14 multiple-choice pen-and-paper questions. Pupils were given 30 minutes to answer all the questions so that the test could be taken during a regularly scheduled teaching hour. The problems and the possible choices were presented on three question sheets and pupils received 3, 4 or 5 points for correct answers, depending on the difficulty level of the questions.[17] There were five answering possibilities with only one correct answer per question, and pupils had to mark their answers on the same

[15] In the 2012/13 school year, there were 2,018 secondary schools in North Rhine-Westphalia (NRW) with 37,451 school classes and a total of 1,295,741 pupils. Of these pupils, 12.26% attended Secondary General School, 23.07% attended Middle School, 18.95% Comprehensive School and 45.72% High School. The share of foreign pupils or pupils with migration background is as follows: *Secondary General School* 57.46%, *Middle School* 39.84%, *Comprehensive School* 46.02%, and *High School* 20.24%.

[16] Schools which responded negatively explained their rejection due to a number of other requests of researchers and lack of time capacities.

[17] There were five questions for three points, five questions for four points and four questions for five points.

sheet. To minimize random answering, one point was deducted for a wrong answer and zero points were given for no answer. To minimize cheating (see Jensen et al. [2002]; Behrman et al. [2015]; Armantier and Boly [2013]), we changed the order of questions for pupils within a class.

The mathematical problems were a compilation of old questions of the *Känguru-Wettbewerb* and differed among school types. We prepared one test for High Schools and another test for Vocational Schools.[18] One test for all school types is not appropriate, as the questions would otherwise be to easy for High School pupils or too difficult for Vocational School pupils. We considerably reduced the length and complexity—particularly the verbal explanations—as many pupils in Vocational Schools have problems understanding lengthy text and lack abstraction capabilities (Retelsdorf and Möller 2008).

To fulfill privacy and data protection requirements, each test and questionnaire received a test identification number, so that pupils did not have to write down their names. This procedure is similar to that of evaluations of learning processes that are regularly carried out in various subjects.

2.2.2 Survey and Selection of Test Incentives

To whom do pupils want to reveal their educational achievement and what type of non-monetary incentives could potentially work in the German school environment? To answer these questions, we conducted a survey before implementing the field experiment in 11 classes of 4 schools with a total of 241 pupils of the same age group. This was a convenience sample gathered through personal contacts. The survey consisted of two parts. On the first page, pupils were asked for three incentives that would motivate them to learn for a test. On the back of the sheet, pupils could mark their choices from a predefined selection. The number of answers was limited to three and pupils were asked to rank the answers (the survey and a complete list of pre-selected incentives can be found in Appendix 2.8). The selection of the reward options in the survey was based on the concepts of Goal Theory,[19] the aspect of work avoidance and social recognition. We categorize these incentives as follows: (i) *work avoidance*, (ii) *mastery*, (iii) *social recognition*, which can be further distinguished in (iiia) *private* and (iiib) *public*, (iv) *consumption* and (v) *curiosity.* Work avoidance incentives lower pupils' "educational inputs" (e.g. homework voucher, bonus points), mastery incentives help pupils to expand their knowledge (e.g. exercise books), social recognition incentives praise pupils' educational achievements (e.g. certificate, teacher praises you in front of the class), consumption incentives

[18] Usually the *Känguru-Wettbewerb* does not differ across school types.

[19] In addition, to set our work in the context of the pedagogical literature we refer to *Goal Theory* which is a widely used concept in pedagogy research. Goal Theory was developed to classify and explain motivation in school (see Ames [1992]) and therefore serves as one source for our survey incentives. The basic idea behind Goal Theory is that there are two main types of motivation: The *Ability Goal Orientation*, i.e., the motivation to be better than classmates and to earn good grades, and *Goal Mastery Orientation*, i.e., the motivation to expand knowledge in one subject and the joy of learning. Goal Theory has been extended by aspects such as *Work Avoidance* (Dowson and McInerney 2001) and *Social Goals* (Urdan and Maehr 1995). They define four types of social goals as sources of motivation: social recognition, social compliance, social solidarity and social care.

are rewards which are unrelated to education (e.g. being allowed to use the mobile phone) and curiosity incentives are to pupils' unknown incentives.

We find that pupils tend to prefer *work-avoidance incentives* and *private social recognition incentives* over *mastery incentives* and *public social recognition incentives*. Overall, the most frequently chosen incentive was extra points for the next exam, which is consistent with the findings of Chevalier et al. (2014), who show that participation in solving quizzes increased between 40% and 60% when the quiz grade counted (2.5% or 5%) toward the course's final grade. The least-favored incentive of our survey was to receive a certificate, which contrasts (at least in stated preferences) with the findings of Jalava et al. (2015)—that girls responded strongly to being rewarded with a certificate—indicating that there might be differences among pupils from different cultural backgrounds. Figure 2.4 in Appendix 2.8 presents the top answers of the survey.

Based on our survey results, we chose to assess the following incentives: (i) medal, (ii) letter of praise and (iii) a choice of incentives where pupils could choose between the medal, the letter, a "no-homework" voucher and a surprise gift. The *homework voucher* could be used once during the semester and exempts pupils from homework in math. The medal is worth about 1 Euro and was awarded in front of the other pupils in the classroom. The *parents-letter* was a pre-formulated letter addressed to parents and signed by the teacher, praising pupils' performance (see Figures 2.5, 2.6 and 2.7 in Appendix 2.8). The *surprise* consisted of the medal plus the parental letter which was not revealed to the students beforehand.

We find small gender differences in the survey. Girls evaluate the *parents letter* slightly higher than boys, whereas boys evaluated the medal higher than girls. In our study, we did not include *Bonus Points* and *Mobile Phone* because teachers are often not allowed to give extra points for the following exam and the use of mobile phones is prohibited in almost all schools. However, *Bonus Points* might be promising to test in future research. For example, Chevalier et al. (2014) have shown that assessment weighting is highly effective among university students.

Finding from Survey: *Pupils prefer work avoidance and private social recognition incentives over mastery goal and public social recognition incentives.*

2.3 Experimental Intervention

The study was conducted in 25 secondary schools with a total of 89 school classes in Bonn, Cologne and Düsseldorf, cities that are located in the federal state of North Rhine-Westphalia, Germany. During February and March 2014, 2.113 pupils in grades 5 and 6 participated; these students were approx. 11 years old, on average, and 43.49% of the participants were female. There might be some selection on the school level, regarding which schools would participate; however, all the pupils of an included class participated. Therefore, we eliminate the potential sample selection bias that might arise with voluntary participation and self-selection of pupils, who are our main subject of interest.

Treatments

We designed the following four treatments to identify pupils' preferred target audience, to analyze the effectiveness of public recognition incentives on academic achievement and to evaluate the power of delegation: the Control Treatment (*Control*), the Letter Treatment (*Letter*), the Medal Treatment (*Medal*) and the Choice Treatment (*Choice*). The test was announced and the preparatory material was distributed one week in advance for all treatments. During the preparation week, teachers did not actively prepare pupils for the test. Teachers answered questions concerning the preparatory exercises only if pupils asked on their own initiative.

Reward Conditions The condition for receiving the reward in all incentivized treatments was an improvement in test grades compared with pupils' last midterm grade. Top-performing pupils who received the highest possible midterm grade received the reward if they did not perform worse. The rationale behind using a relative performance measure is to avoid demotivating low-performing pupils. A criterion-based incentive condition—one in which pupils must score above a predetermined benchmark—might demotivate low-performing pupils because they may believe that the benchmark is not reachable. For example, Jalava et al. (2015) found that girls near and in the lowest decile were demotivated by a high threshold. The grading system of the test was designed such that the highest performing pupil in a class received the highest possible grade and others were graded relative to the top performer. This grading scheme ensured that at least one reward was paid per class. Notably, we focus on the number of test points in our analysis and do *not* consider the difference between the midterm grade and the test grade as a dependent variable in our later analysis.

Control Treatment Pupils in the control group were offered no reward for test performance. For this group, nothing changed from the usual test situation. The test scores of the control group serve as a baseline to estimate the effects of providing non-monetary incentives. The average treatment effect is the difference in the mean test score of each incentivized treatment and the control group.

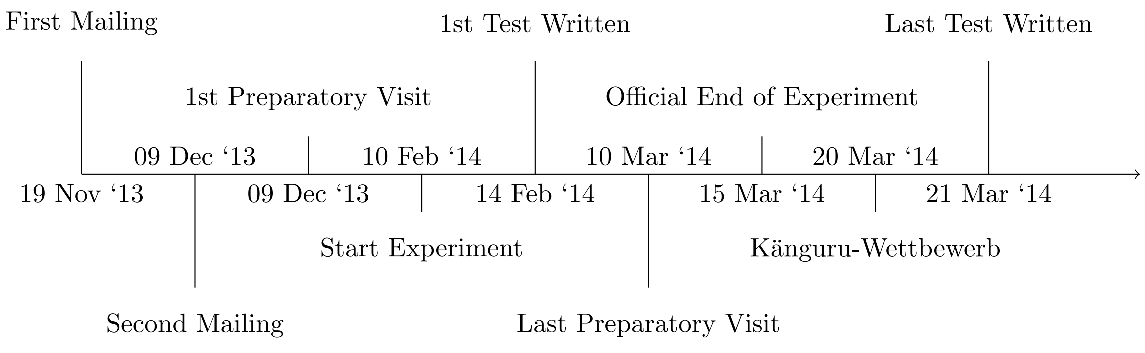
Fixed Treatments (Letter & Medal) In the fixed treatment rewards, one week prior to the test, teachers explained to the pupils that they would earn a reward on the test if they could improve on their last midterm grade. They also explained the reward condition and presented the class with a copy of the rewards. It was further explained that grading the test and thus receiving the reward would take one week at most. One week later, on the test day and shortly before the test, teachers reminded their class of the incentive and explained the reward conditions.

Choice Treatment The Choice Treatment is of main interest as it allows to elicit pupils' preferred signaling target and to evaluate the motivational power of delegation. In contrast to the fixed treatment rewards, pupils assigned to the Choice Treatment were given the choice of incentive beforehand. After announcing the test, each pupil could individually mark one incentive out of four (medal, parental letter, homework voucher and surprise) on a small card. This card was then collected by the teacher and the preparatory materials were distributed. The procedure on the test day remained the same as in the fixed treatments.[20] This treatment was inspired by recent results in real effort experiments, where Mellizo et al. (2014) show that workers that voted to determine their compensation scheme exerted significantly more effort.

Experimental Procedure

We visited the schools one time during the preliminary stage of the experiment. During this meeting, the exact schedule and expiration of the experiment was described and teachers' questions were answered. Each teacher received the instructions (again) in written form near the start of the experiment. In total, two envelopes at different points in time were sent to the teacher. The first envelope was distributed at the beginning of the experiment (February 10, 2014) and contained instructions regarding the announcement of the test, preparatory material for pupils and copies of the rewards to present in front of the class. The teachers communicated the test date to us via email. Two to three days in advance of the test date, teachers received the second envelope containing the actual tests, instructions for the test day and a list in which teachers entered the midterm grades along with the corresponding test-id numbers. Sending the tests in a timely manner was important to reduce the risk that teachers—willingly or unwillingly—prepared pupils. Tests were corrected by the researchers and teachers were asked to answer a questionnaire at the close of the experiment. Figure 2.1 shows a schematic timeline for the experiment.

Figure 2.1: Time-line



Our aim was to maintain a natural exam situation within the classroom. Therefore, the tests took place in regularly scheduled classes in which teachers were free to choose the test date during a predetermined period (February 10th - March 15th). In this manner, teachers could choose a suitable testing week in which no other

[20] Since the number of participating schools was restricted, we did not test the "Voucher Treatment" and "Surprise Treatment". The rationale for choosing the "Medal Treatment" is the comparability to the study by Levitt et al. (2016). The "Letter Treatment" was chosen as this can be easily implemented by teachers and policy makers.

class test was scheduled for which pupils had to study. Furthermore, we had to evaluate the trade-off between a potential loss of control and increased external validity for our results. We opted for the latter, and the experiment was conducted solely by the teachers as we did not want to change the natural class environment and thereby induce experimenter demand effects (see Zizzo [2010] for a discussion of experimenter demand effects). This would have seriously challenged the internal and external validity of our results. Thus, pupils were unaware that the test was part of an experiment.[21]

The test was announced one week in advance and teachers explained the bonus scheme in the event that the class had been assigned to an incentive treatment. In the same lesson, pupils received preparatory questions with attached solutions. Notably, this preparatory material did not prepare pupils with respect to the content of the curriculum but was instead intended to prepare pupils for the (multiple-choice) format of the test. In Section 2.6, we analyze the impact of preparation on pupils' achievement on the test in greater detail. We find that pupils who are significantly more likely to prepare for the test do not perform significantly better on the test. Thus, a difference in pupils' test achievement would not be the result of exerting more effort for test preparation.

The teachers clarified that pupils will be evaluated and graded and that test grades do not count for the school report. They did so in the framework of an evaluation of pupils' achievements that demonstrate their skills during a school year. Hence, the test is low-stakes like the PISA and other standardized comparative tests (i.e. VERA, IGLU, TIMSS). We decided to test low-stakes incentives in a low-stakes testing environment to clearly identify potential (negative) effects of public recognition rewards. In a high-stakes testing environment it could be difficult to identify the incentive effect of low-stakes rewards due to a potentially overlapping incentive effect stemming from the high-stakes testing environment.[22] As the experiment is conducted in pupils' natural learning environment and pupils receive a grade and feedback about their test performance, there are several reasons why pupils should be motivated to excel in this 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 and third, giving grades and feedback on performance allows for social comparison within the classroom (Bursztyn and Jensen 2015). While the former is a "natural" incentive to perform at all and should be represented in all groups, the latter two are additive sources of motivation and the focus of our study. Before the test started, teachers read the following text aloud in the classroom:

[21] According to Zizzo (2010), experimenter demand effects are typically a problem only when they are positively correlated with the true experimental objectives' predictions. In our experiment, this would be the case if an unknown (external) person would have offered pupils a reward (or rewards) for good performance. However, if researchers were never present in the classroom, the pupils' natural environment would remain unchanged because teachers typically try to motivate pupils to increase their efforts in school. Thus we simply changed the way that teachers motivated pupils but not their objective, i.e., improvement in performance.

[22] There is evidence that test performance does not change if the test counts towards the math course grade (Baumert and Demmrich 2001).

> *"The test contains a total of 14 tasks that must be solved within 30 minutes. For each task, there are 4 wrong and 1 correct answers. There are tasks that are worth 3 points for each correct answer, and others that are worth 4 or 5 points. If an incorrect answer is written, 1 point is deducted. If no answer is given, you receive 0 points. Calculators are not allowed, but "scratch paper" for sketches and small calculations are allowed, of course!"*

Pupils then had 30 minutes to answer all the test questions and a questionnaire that was attached to the end of the test. The tests were corrected centrally by the researchers, and the pupils received their rewards one week later.

Randomization procedure Randomization was performed using a classroom-based block randomization design (see Duflo et al. [2007]; Bruhn and McKenzie [2009] regarding the rationale for the use of randomization). As there are at least three classes in almost every school, our treatment assignment procedure ensures that the Control, Choice and at least one of the Fixed Treatments (Medal and/or Letter) was implemented in each school. The Medal and Letter Treatment was implemented simultaneously in schools with more than three classes. Table 2.8 in Appendix 2.8 shows the randomization of treatments over all school types and reports average points by treatment group for the full sample and for boys and girls separately. Table 2.7 in Appendix 2.8 reports the randomization checks for variables we will use as controls in our analysis adjusting for multiple hypothesis testing (see List et al. (2016)). On average, control variables do not differ from the control group at conventional levels of statistical significance, which indicates that the randomization procedure was successful. However, in the treatment groups the share of female teachers seems to be higher and teachers seem to be more experienced. Nevertheless, differences are small and taken into account in our statistical analysis. Subjects in the sample are on average, 11.16 years old and have 0.92 older siblings. 43.49% of the subjects are female and 58.17% speak only German at home, while 37.59% speak another language and 4.24% speak two languages at home. The average midterm grade in mathematics is 2.86 on a scale from 1 to 6, where 1 is the highest and 6 is the lowest grade.

2.4 Data and Descriptive Statistics

We collected data on pupil, teacher and class characteristics. The most important control variable is pupils' last midterm grade which will control for baseline performance in our analysis. In Germany, midterm grades are given on a scale from 1 to 6, where lower numbers represent higher baseline grades. The last midterm grades are reported by teachers and available for almost all pupils. Midterm grades in Germany combine the written and verbal performance of pupils wherein the written part has a larger influence on the final grade; thus, these grades are therefore a good measure of math ability. Importantly, the midterm grades can be treated as pre-determined in our analysis because they were given to pupils before teachers learned about our experiment.

Additional control variables on the individual pupil level are gender, parents' education and a dummy for whether pupils are in grade 5 or 6. 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).

Moreover, we include classroom-level controls: teachers' gender, teachers' working experience and the share of German-speaking pupils within a class. While the literature argues 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, (see Mueller [2013] for a literature review). The influence of teacher's gender on pupils (math) performance has been investigated by Carrell et al. (2010) who find that the professor's gender has little impact on male students but a powerful effect on female students' performance in math. As classes are closed entities with in-part strong peer effects, ethnic and gender composition might have an influence on pupils' performance (see, for example, Jensen and Rasmussen [2011]; Ohinata and Van Ours [2013]). Thus, to control for ethnic and gender composition effects we include the share of German-speaking pupils in the analysis.

We also control for the fact that classes within a school took the test on different days. Therefore, the number of days between the test and the first test written in the respective school is controlled for. Moreover, we distinguish between High Schools and Vocational Schools as our main categorization of interest. The group of High Schools consists of the German Gymnasium whereas Comprehensive, Middle and Secondary Schools belong to the group of Vocational Schools.

Table 2.1 compares the descriptive statistics to the actual data in NRW. Although we cannot claim representativeness of our sample for the school population in NRW, our data are consistent with key school indicators from NRW. We included 2.067 observations in our analysis. 46 observations were dropped because of missing values.[23]

[23] Missing values were the result of incomplete pupil questionnaires. There are 23 missing values for the last midterm grade and 23 for pupils' gender.

Table 2.1: Comparison of Important Indicators: Experiment vs. North Rhine-Westphalia (in percent)

	Experimental Data	*North Rhine-Westphalia*
A. Vocational School		
Proportion Female	44.63	47.39
Proportion Pupil German	57.45	50.11
Class size	26.08	25.37
Proportion Teacher Female	66.59	65.37
B. High School		
Proportion Female	41.99	51.73
Proportion Pupil German	79.74	72.46
Class size	26.83	27.10
Proportion Teacher Female	55.15	59.16

Note: This table presents characteristics of the sample in the experiment by school type and compares it with the same indicators in North Rhine-Westphalia. The cell entries present the percentage shares 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`)

2.5 Results

We first focus on the delegation of choice and report pupils' incentive selection and test whether the target audience varies by pupils' ability or gender. We present descriptive statistics and then estimate multinomial logistic regression controlling for pupil and class characteristics. After this we analyze the effectiveness of providing public recognition incentives on pupils' academic achievement (test scores). First, we present treatment effect estimates—both, pooled and by gender—using negative binomial regression models along with ordinary least square regression models for the pooled Choice Treatment and the Fixed Treatments. We control—among others—for pupils' baseline performance, in order to examine the overall effect of delegation compared to predetermined incentives. Finally, we examine the effectiveness of self-selected incentives for each of the four rewards in the Choice Treatment. Moreover, after having discussed the experimental results, we report further findings on the predictive power of teachers' gender, teachers' working experience and pupils' ability on test performance.

2.5.1 Peers or Parents? On the Selection of Incentives

Classrooms are closed entities with in-part strong peer effects that may determine a child's behavior and the attitude toward learning and effort (Carrell et al. 2009; Kremer et al. 2011; Sacerdote 2011). Giving pupils the possibility to choose their incentive beforehand can shed light on the question to whom pupils primarily want to reveal their educational performance and whether the freedom of choice translates into improved performance. Moreover, as the German system is characterized by different ability tracks, we are interested in answering the question whether pupils who are allocated into different tiers of the education system want to reveal their educational achievement to a different audience?

As early tracking in Germany is not only based on ability, but also on intrinsic motivation and social background, we expect different signaling decisions depending on the school type. Pupils' socio-economic status is likely to differ substantially between school types as families with higher levels of education are more likely to sent their children to High School (even if this is not recommended by the elementary school). This disparity in social background could lead to different effects of signaling decisions because the value for education may differ with the socio-economic background. Sirin (2005) for example shows that children from lower-income families receive less parental involvement than their higher-income classmates which in turn is a signal to lower-income pupils that education is less appreciated by their parents. Thus, if public recognition incentives have different signaling values to the target audience and pupils are free to choose their incentives, we expect differences in the choice of incentive by school type.

Hypothesis 1 *Pupils in Vocational School are more likely to choose an incentive with signaling value to peers (medal) while pupils in High School prefer an incentive with signaling value to parents (letter).*

Choice of incentives Table 2.2 reports pupils' choice over selected incentives by the midterm grade and the school type. Cell entries represent the share of pupils with the same midterm grade—lower numbers represent higher baseline grades—who have chosen the corresponding reward.

The primary incentives of interest are the medal and the letter as this incentives either signal educational achievement to peers or to parents.[24] On the other side, the homework voucher and the surprise should have no clear target audience. Overall, pooling High School and Vocational School students, we observe that the share of pupils who chose the medal or the homework voucher is decreasing in the midterm grade while the share of pupils choosing the surprise is constant over ability levels. Moreover, the share of pupils opting for the letter is clearly increasing moving from high- to low-ability pupils. In other words, high-achieving pupils tend to be more likely to choose an incentive that has a signaling value to their peers and low-ability pupils seem to chose a reward with a signaling value to parents. This is an indication that recognition by parents is of greater importance for the latter.

[24] The medal might also have some signaling value to parents. However, the intended target audience should be the peers as the medal is distributed in front of the peers which are consequently the first target audience to whom pupils would signal their achievements.

Furthermore, the surprise proves to be very popular; approximately one-third of all pupils chose the surprise as their desired reward. One explanation might be the high degree of curiosity among younger children (see Loewenstein [1994] for a psychological perspective of curiosity).

Comparing the choice of incentives separately by school type, we find small and insignificant differences between High Schools and Vocational Schools for high- and middle-performing pupils. However, pupils with midterm grade 4 in High School chose different incentives than their counterparts in Vocational Schools. While 58.82% of those pupils in High School chose the parents letter and only 8.82% chose the surprise, 30.61% of lower-performing pupils in Vocational School chose the parents letter but 35.71% chose the surprise.

Table 2.2: Chosen Incentives by Midterm Grade (in percent)

Midterm Grade	*Medal*	*Letter*	*Voucher*	*Surprise*
Panel A: Pooled (N=676)				
1	32.00	8.00	26.00	34.00
2	19.81	19.32	24.64	36.23
3	20.48	28.11	19.28	32.13
4	15.15	37.88	18.18	28.79
5	10.35	24.11	15.79	31.58
Total	19.53	26.63	21.01	32.84
Panel B: Vocational School (N=399)				
1	36.84	5.26	26.32	31.58
2	20.21	19.15	22.34	38.30
3	17.50	29.38	18.75	34.38
4	14.29	30.61	19.39	35.71
5	14.29	35.71	17.86	32.14
Total	18.05	26.57	20.05	35.34
Panel C: High School (N=277)				
1	29.03	9.68	25.81	35.48
2	19.47	19.47	26.55	34.51
3	25.84	25.84	20.22	28.09
4	17.65	58.82	14.71	8.82
5	0.00	60.00	10.00	30.00
Total	21.66	26.71	22.38	29.24

Note: This table presents the choice of reward of pupils in the Choice Treatment. Cell entries present percentages. Panel A shows the result pooled over school types; panel B presents the choice of pupils in Vocational Schools and panel C the choice of pupils in High Schools. Within the German school system, 1 is the highest and 6 is the lowest possible grade. We do not report on the choice of pupils having the lowest midterm grade because we only have two observations in that group.

Estimating the multinomial logistic regression on the incentive selection distinguished by school types allows us to test whether choices indeed differ across high-

and low-ability performers (Table 2.3). We control for pupil and class characteristics described earlier and the pupils which were not allocated to the Choice Treatment represent the baseline. The variable *midterm grade* is the variable of interest as it measures differences in the incentive choice by pupils' ability.

Pooling over school types, we find that low-ability pupils choose significantly more often the letter sent to the parents (0.093, p = 0.001) and that high-ability pupils choose significantly more often the medal (-0.069, p = 0.082) and the homework voucher (-0.055, p = 0.049).[25] The surprise is also chosen more frequently by high-ability pupils but this finding is not significant (-0.048, p = 0.125). We find the same choices for high- and low-ability pupils if we differentiate between Vocational School and High School which is in contrast to Hypothesis 1 that incentive choice differs by school type. However, the choices described above are significant in High School for the letter (0.148, p = 0.002), homework voucher (-0.086, p = 0.021) and surprise (-0.117, p = 0.036), while the only significant finding in Vocational School is the more frequent choice of the letter by low-ability pupils (0.080, p = 0.014). The latter finding and the fact that high-ability pupils in High Schools tend to be more likely than low-performers to choose the medal and choose significantly more often the letter and surprise shows the strong preference of low-ability pupils to signal their educational achievements to their parents.

The signaling decision of pupils could also differ by gender as boys are usually more competitive than girls and girls usually tend to conform with other expectations. Hence, the medal could be more appealing to boys and the letter to parents more appealing to girls. Table 2.9 in Appendix 2.8 shows the incentive choice for boys and girls separately for school types and for each ability level. In all subgroups, boys tend to be more likely than girls to choose the medal. Looking at the other incentives, it seems that low-performing girls chose more often the parents letter than low-performing boys but this differs for other subgroups. However, all gender differences are not statistically significant. We summarize these findings in our first result:

Result 1 *Pupils with lower grades want to reveal their educational achievement to their parents while high-performing pupils are more likely to chose an incentive with signaling value to peers.*

2.5.2 Incentives and Test Performance

We now turn to the effectiveness of providing public recognition incentives on pupils' test performance. Our primary variable of interest is the number of points scored on the test. Therefore, we will apply negative binomial models and present ordinary least square estimates as robustness check. We first concentrate on the pooled Choice Treatment to analyze the overall effect of delegation compared to predetermined incentives and in a next step, split up the Choice Treatment by chosen incentive. Thereafter, we apply logistic regression to analyze the impact of incentives on preparation time and hence whether treatment effects are driven by an increase in

[25] Estimation results for the homework voucher and the surprise are presented in Table 2.10 in Appendix 2.8

Table 2.3: Multinomial Logit Model of Chosen Incentives

	Pooled		*Vocational School*		*High School*	
A. Medal						
Midterm grade	-0.069*	[0.040]	-0.076	[0.055]	-0.062	[0.057]
Grade 6	-0.230	[0.508]	0.410	[0.719]	-2.343**	[1.026]
Female pupil	-0.186	[0.222]	-0.213	[0.287]	0.084	[0.316]
Books at home						
(11-25)	-0.533	[0.344]	-0.447	[0.371]	-0.363	[0.657]
(26-100)	-0.382	[0.284]	-0.387	[0.313]	-0.578	[0.763]
(101-200)	-0.439	[0.437]	-0.636	[0.767]	-0.419	[0.896]
(201-500)	-0.562	[0.384]	0.495	[0.477]	-1.616	[1.039]
(over 500)	0.086	[0.428]	0.394	[0.606]	-0.421	[1.028]
(Not Reported)	-0.739*	[0.446]	-0.470	[0.649]	-0.912	[0.862]
Teacher experience (years)	0.022	[0.026]	0.034	[0.033]	-0.019	[0.036]
Day difference	0.023	[0.030]	0.007	[0.038]	0.190***	[0.057]
Teacher female	0.173	[0.498]	0.490	[0.710]	-1.181	[1.003]
Unemployment	0.062	[0.072]	0.104	[0.098]	-0.086	[0.089]
Proportion German	-0.952	[1.000]	-2.527	[2.083]	-3.337*	[1.806]
Constant	-0.693	[2.971]	-4.116	[4.264]	14.40**	[6.986]
B. Letter						
Midterm grade	0.093***	[0.029]	0.080**	[0.033]	0.148***	[0.047]
Grade 6	0.462	[0.546]	0.672	[0.745]	-0.736	[1.288]
Female pupil	0.110	[0.208]	0.172	[0.294]	0.063	[0.326]
Books at home						
(11-25)	0.145	[0.289]	0.154	[0.336]	0.037	[0.357]
(26-100)	0.255	[0.335]	0.026	[0.341]	0.368	[0.461]
(101-200)	0.273	[0.428]	0.305	[0.532]	0.181	[0.599]
(201-500)	0.623	[0.552]	0.196	[0.628]	0.478	[0.848]
(over 500)	0.299	[0.523]	0.821	[0.681]	-0.066	[0.731]
(Not Reported)	-0.012	[0.488]	-0.228	[0.581]	0.386	[0.732]
Teacher experience (years)	-0.012	[0.025]	0.002	[0.032]	-0.038	[0.047]
Day difference	0.019	[0.032]	0.006	[0.043]	0.119*	[0.063]
Teacher female	0.660	[0.621]	1.475*	[0.894]	-0.572	[1.187]
Unemployment	0.033	[0.075]	0.010	[0.102]	-0.051	[0.093]
Proportion German	-0.969	[1.002]	-1.706	[1.820]	-2.435	[1.686]
Constant	-5.738	[3.736]	-6.925	[5.286]	3.077	[9.063]

Note: This table presents the results of a multinomial logit model on the choice of incentive of pupils in the Choice Treatment (results of the multinomial logit model for the voucher and surprise are reported in Table 2.10 in Appendix 2.8). The pupils which were not allocated to the Choice Treatment represent the baseline. Midterm grade is the variable of interest, a positive coefficient shows that low performing pupils are more likely to chose the reward as a high midterm grade resembles low performance in the German school system. A negative coefficient shows that high performers are more likely to chose the respective incentive. Covariates: last midterm grade, number of books at home, academic year (grade 5 or 6), gender, teachers' working experience (in years), teachers' gender, day differences between tests and the proportion of German speaking pupils within the class. The number of observation is 2.067 for the pooled specification, 869 for High School and 1.098 for Vocational Schools. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. * p<0.10, ** p<0.05, *** p<0.01.

preparation or an effort effect. Our identification of the average treatment effects on the test score relies on our block randomization strategy. We therefore compare test scores of pupils in the treatment groups to pupils in the control group.[26] Negative binomial models are closely related to the Poisson models that are frequently used for count data, but negative binomial models do not require the restrictive assumption of unitary variance. As our data show a significant degree of overdispersion (Vocational School: $\ln \alpha = -2.004$, p-value < 0.001; High School: $\ln \alpha = -2.636$, p-value $= 0.001$), the negative binomial provides a basis for a more efficient estimation. We control for pupil and class variables described earlier and standard errors are clustered on classroom level—which is the level of randomization. We distinguish between High Schools and Vocational Schools as our main categorization of interest. We estimate the models separately for High Schools and Vocational Schools and allow for school fixed effects.[27] Furthermore, our results are robust to multiple testing—linking equations by seemingly unrelated estimations.[28] This leads us to the following negative binomial model:

$$
\begin{aligned}
E(points_i) \;=\;& m(\beta_0 + \beta_1 \; Treat_i + \beta_2 \; School \; Level_i + \beta_3 \; Midterm_i \\
& + \gamma P_i + \mu C_i + \delta School_i) \qquad (2.1)
\end{aligned}
$$

$m(\cdot)$ is the mean function of the negative binomial model. $points_i$ is the number of points achieved on the test by pupil i, $Treat_i$ indicates the respective treatment, $School\ Level_i$ indicates whether pupils are in grade 5 or grade 6, $Midterm_i$ is the grade in math on the last semester report, P_i is a vector of pupil-level characteristics, C_i a vector of class-level covariates and $School_i$ controls for school fixed effects. As a robustness check, we estimated a linear model (OLS) using the same covariates, and the results change neither in significance nor size.

In Table 2.4, we report on the average treatment effect of the Choice, Medal and Letter Treatments over all school levels.[29] Subsequently, we will report in more detail on the average treatment effects for boys and girls. A special focus is on the Choice Treatment because this is the first study that evaluates the flexibility and freedom of choice on a set of permissible incentives in the educational sector. At first view, the results from Mellizo et al. (2014)—effort increases if workers can determine their compensation scheme—do not seem to extend to the educational sector. Although the coefficients are positive for pupils in the Choice Treatment in all school types, they are small in High Schools (0.091, p = 0.920) and only slightly

[26] Remember, we do *not* consider the difference between the midterm grade and the test grade as a dependent variable.

[27] Note that there has not been a change of teacher between the midterm grade and the test.

[28] Seemingly unrelated estimation combines the parameter estimates, the variance and covariate variances of the separately estimated equations into a robust single parameter-vector and simultaneous variance covariance matrix. The advantage of seemingly unrelated estimations is the robustness to cross-equation correlation and between group heteroskedasticity; consequently, it can overcome the problem of multiple testing.

[29] Kernel density estimations for incentivized and non-incentivized pupils are presented in Appendix 2.8.

larger for those pupils in the Vocational Schools (1.109, p = 0.360) and nonetheless statistically insignificant. However, this null result is interesting when comparing the Choice Treatment with the Fixed Treatments (Medal and Letter) in High School, as treatment effects are significantly negative in the latter. This shows a positive correlation between pupils motivation and delegation.[30]

Comparing school types, predetermined incentives seem to work in opposite directions. In the Medal and Letter Treatment there are no significant differences in Vocational Schools but significant negative effects in High Schools (Medal: -2.006, p = 0.033; Letter: -2.586, p = 0.058).[31] In High Schools, we can reject that the insignificant positive estimate in the Choice Treatment is not equal to the significant negative effects of the Medal (p = 0.066) and Letter Treatment (p = 0.093).

Result 2 *Educational achievement decreases for pupils in High School if public recognitional incentives are predetermined but not if they can be freely chosen.*

[30] Power calculations show that the minimal detectable effect size with the present sample size and randomization strategy is 0.157-0.178 standard deviations (depending on whether alpha is 0.05 or 0.10). Even if the effects turn out to be significant in larger samples, the interpretation of our results would not change.

[31] Ordinary least square estimation on grade improvement—difference between midterm grade and grade in test—shows similar results. The Choice and Medal Treatments in Vocational Schools have (insignificant) positive coefficients—pupils received better grades in the test compared to their midterm grade—and the Letter Treatment has a (insignificant) negative coefficient. Treatments in High School have negative and significant effects in the Medal and Letter Treatment. Overall, the percentage of pupils who improved their grade and hence received a reward in the incentivized treatments, is as follows: (i) Vocational School: Control 25.00%, Choice 25.44%, Medal 33.50%, Letter 17.00% (ii) High School: Control 29.88%, Choice 19.35%, Medal 16.96%, Letter 18.50%.

Table 2.4: Treatment Effects

	OLS				Negative Binomial			
	Vocational School		*High School*		*Vocational School*		*High School*	
Treatments								
Choice	1.109	[1.128]	0.087	[0.869]	1.109	[1.211]	0.091	[0.907]
Medal	0.597	[0.991]	-1.709*	[0.918]	0.362	[1.143]	-2.006**	[0.943]
Letter	0.941	[1.192]	-2.262	[1.523]	0.867	[1.219]	-2.586*	[1.364]
Controls								
Pupil Covariates	Yes		Yes		Yes		Yes	
Class Covariates	Yes		Yes		Yes		Yes	
School FE	Yes		Yes		Yes		Yes	
N	1198		869		1198		869	

Note: This table compares the result of a linear and negative binomial regression separately for High Schools and Vocational Schools including school fixed effects. Dependent variable: points in test. Covariates: last midterm grade, gender, number of books at home, academic year (grade 5 or 6), teachers' working experience (in years), teachers' gender, day differences between tests and the proportion of German speaking pupils within the class. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. Robustness checks with multiple testing—seemingly unrelated regressions—show similar results. * p<0.10, ** p<0.05, *** p<0.01.

As discussed in the Introduction, stereotype-threats and non-conformity to role behavior may cause girls not to excel under incentives that emphasize personal achievement in mathematics. Table 2.5 reports average treatment effects for boys and girls by school type controlling for pupil and class covariates as well as school fixed effects. We find significant gender differences in the reaction to incentives for pupils in the Letter Treatment in High Schools. The test performance of boys decreases significantly in the Letter Treatment (-3.661, p = 0.005), whereas this decrease is not statistically significant for girls. In the Vocational School sample, the coefficients of recognitional incentives have positive signs but are insignificant at conventional levels for boys and girls. We summarize this in our third result:

Result 3 *Public recognition incentives have no heterogeneous gender effects in Vocational Schools. The letter of praise sent to parents is detrimental to the test performance of High School boys.*

Table 2.5: Treatment Effects by Gender

	Vocational School		*High School*	
Males				
Choice	1.129	[1.347]	-0.186	[1.317]
Medal	0.055	[1.112]	-1.073	[1.201]
Letter	0.924	[1.255]	-3.661***	[1.305]
N	665		504	
Females				
Choice	1.209	[1.326]	0.704	[1.303]
Medal	1.043	[1.546]	-2.707	[1.759]
Letter	0.929	[1.380]	-1.136	[2.011]
N	533		365	
Controls				
Pupil Covariates	Yes		Yes	
Class Covariates	Yes		Yes	
School FE	Yes		Yes	

Note: This table reports the result of a negative binomial regression separately for boys and girls and for High Schools and Vocational Schools including school fixed effects. Dependent variable: points in test. Covariates: last midterm grade, number of books at home, academic year (grade 5 or 6), teachers' working experience (in years), teachers' gender, day differences between tests, unemployment rate of the school district and the proportion of German speaking pupils within the class. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. * p<0.10, ** p<0.05, *** p<0.01.

Chosen incentive and test performance We can now examine the correlations between each of the chosen incentives and pupils' test performance. Table 2.6 presents the estimates of the average effects on test performance for each reward in the Choice Treatment as well as the two Fixed Treatments. In Vocational Schools, we find a large positive and significant effect for those pupils who chose the letter (2.640, p = 0.085) which is in line with our findings on the incentive choice. In High School, we find no statistically significant relationship for any of the chosen incentives. However, the difference between the positive coefficient of the "Chosen Medal" and the negative effect of the "Fixed Medal" is significant (chi2 = 5.50, p = 0.019).

Result 4 *Pupils in Vocational Schools who want to signal their educational achievements to parents can significantly increase their test performance. When free to choose the medal in High School, the crowding out disappears.*

Table 2.6: Chosen Incentives and Points in Test

	Vocational School		*High School*	
Treatments				
Medal_Chosen	1.745	[1.887]	0.545	[0.913]
Letter_Chosen	2.640*	[1.532]	-0.446	[1.552]
Voucher_Chosen	-0.650	[1.572]	-0.970	[1.007]
Surprise_Chosen	0.873	[1.439]	1.160	[1.311]
Medal_Fixed	0.419	[1.123]	-1.932**	[0.943]
Letter_Fixed	0.973	[1.122]	-2.547*	[1.358]
Controls				
Pupil Covariates	Yes		Yes	
Class Covariates	Yes		Yes	
School FE	Yes		Yes	
N	1198		869	

Note: This table reports the result of a negative binomial regression for each incentive in the Choice Treatment and the incentives in the Fixed Treatments separately for High Schools and Vocational Schools including school fixed effects. Dependent variable: points in test. Covariates: last midterm grade, gender, number of books at home, academic year (grade 5 or 6), teachers' working experience (in years), teachers' gender, day differences between tests and the proportion of German speaking pupils within the class. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. * p<0.10, ** p<0.05, *** p<0.01.

Our results can be linked to the study by Bursztyn and Jensen (2015). As the medal in our study has a strong signaling value within the peer group, pupils' performances in the (fixed) Medal Treatment can be compared to pupils' performance in the natural experiment of Bursztyn and Jensen (2015). The authors find that high-ability pupils' performance worsened, whereas low-ability pupils improved their performance. This indicates that high-performers try to mimic low performance and low-performers try to mimic high performance. However, in contrast to Bursztyn and Jensen (2015), we do not find that low-performers in the Medal Treatment significantly increase their performance or that high-performers significantly decrease their performance (see Table 2.12 in Appendix 2.8). Our findings are more in line with the theoretical model of Austen-Smith and Fryer (2005) who show that in the single-audience case there is a separating equilibrium. Notably, we find that middle-ability pupils significantly decrease their performance—thus mimicking low-ability pupils instead of high-ability pupils.

What might lead to the crowding out of motivation? Notably, treatment effects are significantly negative in the (fixed) Medal and Letter Treatment in High Schools. One reading of this result is that the incentives provided are too low-powered for pupils from higher socio-economic backgrounds. In the pupil questionnaire, we asked pupils how much the respective reward motivated them on a 1 (not at all) to 5 (very much) scale to verify whether the external rewards differ in their motivational power. We find that pupils in High Schools state a lower motivation than pupils in Vocational Schools in the Choice (3.587 v. 3.340, $p = 0.002$), Medal (3.473 v. 2.971, $p < 0.001$) and Letter Treatment (3.526 v. 2.714, $p < 0.001$). It is also conceivable that pupils in High School are more likely to be monetarily rewarded by parents for good grades. In Germany, parents from the highest tercile of the income distribution spent a monthly average of 160 Euro, middle tercile: 101 Euro and bottom tercile: 74 Euro on the school education of their children.[32] However, this cannot explain a *decrease* in performance in High Schools instead of just a smaller increase, as pupils in High Schools on average report that they are more than “not at all” motivated by the rewards.

Another explanation might be that reputational motivation differs with socio-economic background. Bénabou and Tirole (2006) assume that an agents' pro-social or antisocial behavior reflects an endogenous and unobservable mix of *intrinsic*, *extrinsic* and *reputational* motivations, whereas in our context prosocial behavior is the effort in school (see also Ariely et al. [2009] on image motivation and incentives). The authors show that extrinsic rewards can result in a crowding out of the reputational value of good deeds because they create doubts about intentions, i.e. to what extent was performance increased for the incentive rather than for yourself?

There is evidence that the value for education and hence the reputation associated with academic achievements differs with socio-economic background. In a meta-study, Sirin [2005] shows that children from lower income families receive less parental attention in educational matters than their higher income classmates. According to Dwyer and Hecht [1992], one reason for low parental involvement in the education of their children might be a negative parental attitude. Parents who were never very successful in school or for whom school was a traumatic experience might not send a positive message to their children regarding the importance of education. Hence, an extrinsic non-monetary reward given by teachers could give pupils from lower income families the recognition that they do not receive otherwise or signal to parents that they should praise their child. In terms of the model developed by Bénabou and Tirole [2006], (positive) extrinsic motivation exceeds the amount of (negative) reputational motivation.

[32] These figures are available at `http://www.vodafone-stiftung.de/ideen_foerdern_publikationen.html?&tx_newsjson_pi1[showUid]=30&cHash=e2270fb5104907e5c3be9121af72e237` (accessed November 12, 2015).

2.6 Further Explanations and Results

Do treatment effects result due to increased test preparation or greater effort in the test?

It is often difficult to disentangle whether improvements in educational outcomes are the result of increased efforts in studying the subject or the result of higher effort in solving test questions on the test day. The experimental design by Levitt et al. [2016]—incentives are announced immediately before the test with no advance notice—is one of the few studies that isolates the effort effect by not giving time to prepare for the test. These authors can therefore attribute the incentive effects to greater short-run effort. Providing incentives without advance notice was not possible in our study because our aim was to analyze the effect of giving pupils the flexibility and freedom to choose an incentive. The choice of the preparatory material—old versions of the *Känguru-Wettbewerb*—nevertheless allows us to isolate the effects of short-run effort. It is unlikely that pupils gained knowledge of the subject matter by solving the preparatory material. The material was designed *not* to prepare pupils with respect to the content of the curriculum but to familiarize them with the multiple-choice testing format.

As we have previously shown, we find heterogeneous effects of the incentivized treatments on performance. In the pupil questionnaire, we asked pupils to state whether they prepared for the test using the provided material. Accordingly—if there is a learning effect—these incentivized pupils should also have prepared more or less often than pupils in the control treatment.

Table 2.14 in Appendix 2.8 presents estimates of a logistic regression. The dependent dummy variable is whether pupils (self-reported) prepared for the test. We control for pupils' gender, school level, midterm grade, whether pupils like math (measured on a 1–5 scale) and include school fixed effects.

We find that in all school types, pupils' willingness to prepare for the test in the incentivized treatments is higher than for pupils in the Control Group. The only exception are fifth graders of Vocational Schools in the Letter Treatment (-0.095, $p = 0.301$), which are less likely to prepare for the test than the Control Group. Overall, we find that, in particular, pupils in the Choice Treatment significantly increased the time they have spent on preparation. The results are significant for pupils (grade 5 and 6) in High Schools in the Choice (grade 5: 0.136, $p = 0.007$; grade 6: 0.173, $p = 0.003$) and Letter Treatment (grade 5: 0.075, $p = 0.045$; grade 6: 0.166, $p < 0.001$) and for fifth graders of Vocational Schools in the Choice (0.208, $p = 0.011$) and Medal Treatment (0.186, $p = 0.067$).

We can now compare pupils' willingness to prepare for the test with their actual test performance. Those pupils whose willingness to prepare for the test is positive do not gain significantly more points in the test. Furthermore, those who significantly improved or decreased performance—compared to the control group—have not prepared significantly more or less for the test. These results are an indicator that there is indeed no direct link between test preparation and test performance.

Further Results

A large part of the research in the economics of education involves the effects of educational inputs, such as teacher gender, teacher quality, and/or school resources on pupils' achievement. However, no consensus has been reached regarding how these factors influence student' performance (see Hanushek [1986]; Card and Krueger [1992]; Hoxby [2000]; Rivkin et al. [2005]). We now examine the correlations of some input factors and achievement to see how our sample compares with previous studies. Table 2.15 in Appendix 2.8 shows the coefficients of teachers' working experience, parents' educational background, the gender of teachers and midterm grades on pupils' achievement for the whole sample as well as gender and pupils' ability.

Socio-Economic status and *Books at Home* To further investigate the role of the socio-economic and foremost educational background of the parents, we include *Books at Home* as an explanatory variable. As expected, we find that the number of books is positively correlated with pupils' school achievement in both High Schools and Vocational Schools. The effect seems to be strongest for pupils in High Schools whose parents have more than 200 books at home (201–500 books: 5.065, p = 0.002; over 500 books: 5.095, p = 0.003, Table 2.15 in Appendix 2.8). Furthermore, there is a higher correlation between the education of the household and school achievement for boys than for girls. Analyzing the performance of German elementary pupils in the *Trends in International Mathematics and Science Study 2011* (TIMSS), Bos et al. (2012) find that fourth graders whose families have more than 100 books at home are one year ahead in mathematical skills in comparison with fourth graders who report that their families have fewer than 100 books at home. In our sample, pupils in High Schools responded to have 101–200 books at home, whereas the modal response for pupils in Vocational Schools was fewer than 100 books at home. Given the results of Bos et al. (2012) and the finding in the pupil questionnaire, fifth graders in High School seem to be one year ahead of fifth graders in Vocational School which would be in line with the tracking system in Germany.

Ability Performance differences are driven not only by ability, but also by the amount of intrinsic motivation for the matter at issue. However, the previous literature has shown that extrinsic incentives tend to crowd out motivation for intrinsically motivated tasks (Frey 1994; Gneezy and Rustichini 2000; Frey and Jegen 2001). By asking pupils about their affinity for mathematics on a 1 (not at all) to 5 (very much) scale, we can approximate whether low- and high-performing pupils differ in their intrinsic motivation. We find that high-performers have a significantly higher affinity toward mathematics (3.984) than low-performers (3.150). Hence, providing extrinsic non-monetary incentives to pupils might lead to a poorer test performance for high-ability pupils if a potentially stronger internal motivation gets crowded out. Conversely, low-performing pupils—who lack internal motivation—might benefit by being extrinsically incentivized. Based on externally given midterm grades, we group pupils into *high*-, *middle*- and *low*-ability pupils. High-ability pupils refers to those with a midterm grade of 1 or 2; middle-ability pupils have a midterm grade of 3 and low-ability pupils are those with a midterm grade of 4, 5 or 6. The groups are of approximately equal size.

Table 2.12 in Appendix 2.8 reports the average treatment effects by ability. We find differences between low- and high-ability pupils and differences between pupils at High Schools and Vocational Schools. Motivation is crowded out for low-ability pupils in High Schools in the Medal (-4.480, p = 0.013) and Letter Treatment (-5.672, p = 0.011). By contrast, high-performers in High School do not seem to respond to the rewards in the Choice (0.595, p = 0.540), Medal (-0.240, p = 0.852) and Letter Treatment categories (-2.093, p = 0.262). In Vocational Schools, public recognition incentives enhance test performance for both low-ability and high-ability pupils but decreases test performance for medium-ability pupils. The effects are significant for high-ability pupils in the Letter Treatment (2.791, p = 0.011) and for middle-ability performing pupils in the Medal Treatment (-4.201, p = 0.009).

Our results can be compared to those of Leuven et al. (2010), who find that monetary incentives increase academic performance for the most able students but decrease performance for low-ability students. We find similar results for high-performers in Vocational Schools and low-performers in High Schools. There are at least two mechanisms that can explain the results of Leuven et al. (2010): a pure "crowding out" effect and a "resignation" ("I won't make it in any case") effect. We find similar results although we use non-monetary rewards and different rewarding conditions. Students in the study of Leuven et al. (2010) had to pass all first-year requirements within one year according to a fixed (i.e., non-personalized) threshold. By using a relative rewarding scheme—pupils in our study had to improve relative to their past performance—we reduce or eliminate the "resignation effect". Overall, we find that average treatment effects are positive and highest for high-ability pupils in Vocational Schools.

We can further analyze whether there are gender differences for low-, middle- and high-achieving pupils. Table 2.13 in Appendix 2.8 reports negative binomial estimates differentiated by ability and gender. We find pronounced and large gender differences for low-achieving pupils. Boys do not significantly respond to any type of incentive in High Schools and Vocational Schools. By contrast, intrinsic motivation is crowded out for girls in High Schools in the Choice (-4.727, p = 0.062) and Letter Treatments (-6.208, p = 0.062). In Vocational Schools, girls motivation is increased in the Medal Treatment (4.098, p = 0.029).

Teachers' working experience

Rivkin et al. (2005) show that mathematics teachers in their first year and—to a lesser extent—second- and third-year teachers perform significantly worse than more experienced teachers. There may be some additional gains to experience in the subsequent year or two, but the estimated benefits are small and not statistically significant in both mathematics and reading (see also Harris and Sass [2011]). In line with Rivkin et al. (2005), we find that teachers' experience is correlated with higher achievement in Vocational Schools (0.090, p = 0.015) and High Schools (0.058, p = 0.071). Boys in High Schools (0.077, p = 0.024) and girls in Vocational Schools (0.150, p = 0.002) achieve significantly higher test scores with a more experienced teacher, although there is no significant effect on girls in High Schools

(0.019, p = 0.701) and boys in Vocational Schools (0.061, p = 0.106). Furthermore, low-ability pupils in Vocational Schools (0.181, $p < 0.001$), in particular, perform better than those with an inexperienced teacher. In addition, low-ability pupils in High Schools (0.165, p = 0.037) have better test scores with a teacher who has more experience. In all school types, there is no significant correlation for high-ability pupils. The most common channels in the literature that may explain these correlations are that (i) experienced teachers are better able to use teaching strategies that respond to students' needs and learning styles, (ii) experienced teachers focus more on low-ability pupils and (iii) experienced teachers can better handle disturbances in class.

Teachers' gender

The results of the influence of teacher's gender are mixed so far. Carrell et al. (2010) report that the gender of the professor has little impact on male students' performance in math but a powerful effect on female students' performance, whereas Antecol et al. (2015) find that in primary school, female students who were assigned to a female teacher suffered from lower math test scores at the end of the academic year. Our experimental data support the findings of Antecol et al. (2015). We find that having a female teacher lowers (non-significantly) test scores for girls in High School (-0.387, p = 0.736) and Vocational School (-2.909, p = 0.026). In contrast to Carrell et al. (2010) and Antecol et al. (2015), we find a significant correlation for boys in High School (-3.109, $p < 0.001$), although the correlation for boys in Vocational School (0.429, p = 0.650) is insignificant.

2.7 Conclusion

In this paper, we investigate in a field experiment to whom pupils prefer to reveal their educational achievements by delegating the choice over public recognition incentives which vary with respect to the target audience. We then compare the educational achievements in a mathematical test for pupils offered predetermined rewards for self-improvement to pupils who are free to chose their incentive scheme. This is important to better tailor and increase the effectiveness of non-monetary incentives in the educational system. The selection of rewards is motivated by a survey conducted in the run-up to the experiment asking pupils about their preferences over seventeen pre-selected rewards. We finally tested the following four incentives in the field: (i) medal, (ii) letter of praise, (iii) "no-homework" voucher and (iv) surprise gift.

The experimental design allows to clearly analyze treatment effects, as pupils did not know that they were part of an experiment. We maintain a natural examination situation within the classroom by having the students take the test during a regularly scheduled math lesson and letting teachers conduct the experiment by themselves.[33]

Our general findings suggest that low-ability pupils prefer to signal their academic achievements to parents while high-ability pupils tend to opt for their peers.

[33] Counter arguments for this kind of design might be a potential loss of control. However, we believe that teachers had no incentive to not follow our instructions.

Moreover, we show that public recognition incentives can be a potentially cost-effective way to increase achievement in all school types if they can be freely selected. On predetermined incentives, we find differences by school types. Pupils in Vocational School (pupils most likely from lower socio-economic families) tend to increase their performance while intrinsic motivation is significantly crowded-out in High Schools (pupils most likely from higher socio-economic families). Thus, endogenous selected rewards mitigate the negative effects of extrinsically determined rewards in High Schools. We also find suggestive evidence that empowerment of pupils is beneficial to increase learning inputs; pupils who were free to choose their incentive reported more often to have learned for the test.

A limitation of the experiment, as in most experiments, is that we can only learn the impact of treatments on the population studied, which is a broad—but not representative—sample of the population of pupils in Germany. However, we shed light on pupils' preferred target audience and the effectiveness of public recognition incentives in schools. We conclude that these kind of incentive must be carefully designed and that pupils' socio-economic background—as measured by the school type—must be taken into account.

The applicability of our incentives was confirmed by teachers. Overall, 44.31% of teachers are planning to use at least one incentive in the future. However, while incentives are well received by teachers at Vocational Schools, only about 14% of High School teachers plan to use incentives in the future.[34]

It is important to analyze to whom pupils want to reveal their educational achievement and in particular the choice of the target audience of different ability level, to better inform policy makers and to better tailor large scale interventions in the educational sector. As our results show, the target audience differs by pupils' ability and more importantly, letting pupils participate in the educational process seems to be a promising mechanism to increase educational outcomes. Moreover, it remains for future research to analyze the effectiveness of public recognition incentives in the long term and to test the working of a *"Fixed Voucher"* and *"Fixed Surprise"* treatment, in addition to testing the remaining incentives suggested in our survey in Subsection 2.2.2. It would be also interesting to further investigate the selection of the target audience and the impact of public recognitional incentives on pupils with different cultural backgrounds. Finally, more research is required on identifying potential non-monetary incentives for teachers and to analyze the interplay of all big stakeholders of the educational production function: peers, parents and teachers.[35]

[34] The share of teachers planning to use future incentives by school types is roughly as follows: Secondary General School 66%; Middle School 56%; Comprehensive School 75%; High School 14%.

[35] Having informally asked teachers about their preferences, work avoidance (e.g., somebody else corrects tests) and public recognition incentives seem to be the most promising types of non-monetary incentives for teachers.

2.8 Appendix

Randomization Tables

Table 2.7: 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. Choice	0.1190	0.0233*	0.3537	0.5600	0.4667
	Control vs. Medal	0.0551	0.3383	0.9920	1.0000	1.0000
	Control vs. Letter	0.0159	0.7777	0.9987	1.0000	1.0000
Month of Birth	Control vs. Choice	0.1241	0.6180	0.9990	1.0000	1.0000
	Control vs. Medal	0.3202	0.2633	0.9753	1.0000	1.0000
	Control vs. Letter	0.4766	0.0940*	0.7760	1.0000	1.0000
Num. Older Sib.	Control vs. Choice	0.0041	0.9497	0.9987	1.0000	1.0000
	Control vs. Medal	0.1488	0.0310**	0.4217	0.7440	0.5890
	Control vs. Letter	0.0675	0.3543	0.9903	1.0000	1.0000
Female Pupil	Control vs. Choice	0.0192	0.4813	0.9943	1.0000	1.0000
	Control vs. Medal	0.0254	0.4150	0.9907	1.0000	1.0000
	Control vs. Letter	0.0016	0.9643	0.9643	1.0000	0.9643
Language German	Control vs. Choice	0.0059	0.8273	0.9963	1.0000	1.0000
	Control vs. Medal	0.0343	0.2753	0.9683	1.0000	1.0000
	Control vs. Letter	0.0145	0.6507	0.9960	1.0000	1.0000
Teacher Female	Control vs. Choice	0.1123	0.0003***	0.0003***	0.0080***	0.0080***
	Control vs. Medal	0.0652	0.0423**	0.5073	1.0000	0.7620
	Control vs. Letter	0.0922	0.0010***	0.0133**	0.0240**	0.0210**
Teacher Exp.	Control vs. Choice	2.4663	0.0003***	0.0003***	0.0080***	0.0073***
	Control vs. Medal	5.2002	0.0003***	0.0003***	0.0080***	0.0077***
	Control vs. Letter	0.9000	0.1673	0.9140	1.0000	1.0000
Books Home	Control vs. Choice	0.1359	0.0953*	0.7583	1.0000	1.0000
	Control vs. Medal	0.0514	0.5593	0.9963	1.0000	1.0000
	Control vs. Letter	0.1298	0.1770	0.9137	1.0000	1.0000

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. * p<0.10, ** p<0.05, *** p<0.01.

Table 2.8: Treatment Randomization: Average Test Scores

	Control	*Choice*	*Medal*	*Letter*
Vocational School				
Full Sample				
N individuals	366	404	207	253
Test Score	8.945	10.129	11.275	11.245
	(10.650)	(11.719)	(11.823)	(11.580)
Standardized Test Score	-0.247	-0.141	-0.041	-0.039
	(0.953)	(1.049)	(1.058)	(1.036)
Boys				
N individuals	202	209	115	150
Test Score	10.203	10.861	11.461	12.560
	(10.474)	(11.568)	(12.125)	(12.549)
Standardized Test Score	-0.135	-0.076	-0.022	0.076
	(0.937)	(1.0354)	(1.085)	(1.123)
Girls				
N individuals	160	193	89	102
Test Score	7.519	9.368	10.865	9.157
	(10.758)	(11.914)	(11.562)	(9.635)
Standardized Test Score	-0.375	-0.209	-0.075	-0.228
	(0.963)	(1.066)	(1.035)	(0.862)
High School				
Full Sample				
N individuals	242	279	189	173
Test Score	14.888	14.147	12.206	13.486
	(9.689)	(10.299)	(10.755)	(11.269)
Standardized Test Score	0.285	0.218	0.045	0.159
	(0.867)	(0.922)	(0.963)	(1.009)
Boys				
N individuals	137	164	116	88
Test Score	16.431	14.793	13.543	13.352
	(10.566)	(10.645)	(11.098)	(10.187)
Standardized Test Score	0.423	0.276	0.164	0.147
	(0.946)	(0.953)	(0.993)	(0.912)
Girls				
N individuals	99	115	70	81
Test Score	12.929	13.226	9.971	13.580
	(8.144)	(9.758)	(9.575)	(12.603)
Standardized Test Score	0.109	0.136	-0.155	0.168
	(0.729)	(0.873)	(0.857)	(1.128)

Note: The table displays the descriptive statistics of test scores and the number of students in each of the treatment groups and the control group. Both average points scored on the test and standardized test scores (with mean 0 and standard deviation 1). Standard errors are displayed in parentheses. In our final analysis, we included 2.067 observations. 46 observations were dropped because missing values. There are 23 missing values for the last midterm grade and 23 for pupils' gender.

Selection of incentives

Table 2.9: Chosen Incentive by Gender and School Type (in percent)

	Medal	*Letter*	*Voucher*	*Surprise*
Vocational Schools (Fisher's exact = 0.136)				
Male (N=208)	21.63	27.88	19.23	31.25
Female (N=193)	14.51	24.87	20.21	40.41
High Schools (Fisher's exact = 0.744)				
Male (N= 163)	23.31	24.54	22.70	29.45
Female (N=114)	19.30	29.82	21.93	28.95
High Performers (Fisher's exact = 0.212)				
Male (N= 144)	26.39	17.36	25.00	31.25
Female (N= 113)	16.81	16.81	24.78	41.59
Middle Performers (Fisher's exact = 0.711)				
Male (N= 125)	22.40	29.60	18.40	29.60
Female (N= 122)	18.03	27.05	19.67	35.25
Low Performers (Fisher's exact = 0.584)				
Male (N=102)	16.67	35.29	17.65	30.39
Female (N= 69)	10.14	43.48	17.39	28.99
All School types (Fisher's exact = 0.185)				
Male (N=371)	22.37	26.42	20.75	30.46
Female (N=307)	16.29	26.71	20.85	36.16

Note: This table reports the percentage share of pupils' choice by gender separately for school types and ability levels. Fisher' s exact test reports on the difference in the proportions between boys and girls.

Table 2.10: Multinomial Logit Model of Chosen Incentives Cont.

	Pooled		*Vocational School*		*High School*	
C. Voucher						
Midterm grade	-0.055**	[0.028]	-0.029	[0.039]	-0.086**	[0.037]
Grade 6	0.078	[0.586]	-0.346	[0.759]	-0.264	[1.077]
Female pupil	0.199	[0.232]	0.287	[0.247]	-0.086	[0.397]
Books at home						
(11-25)	-0.287	[0.282]	-0.374	[0.289]	0.201	[0.530]
(26-100)	-0.347	[0.337]	-0.580*	[0.336]	0.424	[0.781]
(101-200)	-0.436	[0.367]	-0.617	[0.379]	0.484	[0.724]
(201-500)	-0.516	[0.495]	-1.405	[0.918]	0.281	[0.810]
(over 500)	-0.380	[0.451]	0.409	[0.590]	0.083	[0.594]
(Not Reported)	-0.340	[0.499]	-1.299*	[0.690]	1.001	[0.942]
Teacher experience (years)	0.018	[0.028]	0.072*	[0.038]	-0.039	[0.040]
Day difference	0.038	[0.029]	0.005	[0.042]	0.125*	[0.069]
Teacher female	-0.317	[0.608]	-0.218	[0.864]	-0.778	[0.869]
Unemployment	-0.030	[0.071]	-0.035	[0.113]	-0.123	[0.111]
Proportion German	0.390	[1.134]	0.260	[1.896]	-1.241	[2.319]
Constant	-2.483	[3.632]	-0.794	[4.698]	1.768	[7.382]
D. Surprise						
Midterm grade	-0.048	[0.031]	-0.031	[0.031]	-0.117**	[0.056]
Grade 6	0.536	[0.513]	1.470**	[0.744]	-1.274	[1.253]
Female pupil	0.313*	[0.177]	0.590***	[0.200]	-0.005	[0.280]
Books at home						
(11-25)	-0.225	[0.272]	-0.127	[0.294]	0.321	[0.508]
(26-100)	-0.231	[0.347]	-0.040	[0.326]	-0.103	[0.859]
(101-200)	0.008	[0.365]	0.047	[0.328]	0.437	[0.917]
(201-500)	0.294	[0.418]	0.515	[0.450]	0.262	[0.950]
(over 500)	-0.274	[0.538]	0.694	[0.680]	-0.559	[0.997]
(Not Reported)	-0.774*	[0.415]	-0.182	[0.422]	-13.92***	[0.904]
Teacher experience (years)	0.015	[0.024]	0.025	[0.033]	-0.009	[0.038]
Day difference	0.016	[0.030]	0.021	[0.047]	0.176**	[0.073]
Teacher female	0.821	[0.546]	1.087	[0.842]	-0.327	[0.940]
Unemployment	0.072	[0.072]	0.194	[0.125]	-0.132	[0.108]
Proportion German	-1.144	[1.108]	-2.480	[1.747]	-3.760*	[2.272]
Constant	-5.258	[3.304]	-11.77**	[5.179]	8.948	[8.798]
N	2067		1198		869	

Note: This table presents the results of a multinomial logit model on the choice of incentive of pupils in the Choice Treatment. The pupils which were not allocated to the Choice Treatment represent the baseline. Midterm grade is the variable of interest, a positive coefficient shows that low performing pupils are more likely to chose the reward as a high midterm grade resembles low performance in the German school system. A negative coefficient shows that high performers are more likely to chose the respective incentive. Covariates: last midterm grade, number of books at home, academic year (grade 5 or 6), gender, teachers' working experience (in years), teachers' gender, day differences between tests and the proportion of German speaking pupils within the class.The number of observation is 2.067 for the pooled specification, 869 for High School and 1.098 for Vocational Schools. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. * p<0.10, ** p<0.05, *** p<0.01.

Robustness Checks

Table 2.11: Robustness Check—Treatment Effects without Covariates

	Vocational School		*High School*	
Treatments				
Choice	2.064**	[1.000]	-0.941	[1.160]
Medal	1.188	[1.027]	-3.511**	[1.623]
Letter	1.344	[1.058]	-1.019	[1.709]
Controls				
Pupil Covariates	No		No	
Class Covariates	No		No	
School FE	Yes		Yes	
N	1230		883	

Note: This table reports the result of a negative binomial regression without covariates separately for High Schools and Vocational School including school fixed effects. Standard errors are reported in parentheses and clustered on classroom-level. Dependent variable: points in test. The number of clusters is 36 in High Schools and 53 in Vocational Schools. Results are robust to multiple testing (seemingly unrelated estimation). * p<0.10, ** p<0.05, *** p<0.01.

Treatment effects by ability (and gender)

Table 2.12: Treatment Effects by Pupils' Midterm Grade

	Vocational School		*High School*	
Low Performing Pupils				
Choice	0.421	[1.119]	-2.969	[1.907]
Medal	0.572	[1.326]	-4.480**	[1.796]
Letter	0.512	[1.387]	-5.672**	[2.229]
N	408		144	
Middle Performing Pupils				
Choice	0.033	[1.911]	0.831	[1.794]
Medal	-4.201***	[1.617]	-3.069*	[1.564]
Letter	-1.407	[1.779]	-3.803**	[1.910]
N	428		288	
High Performing Pupils				
Choice	1.579	[1.267]	0.595	[0.970]
Medal	0.979	[1.208]	-0.240	[1.284]
Letter	2.791**	[1.103]	-2.093	[1.868]
N	362		437	
Controls				
Pupil Covariates	Yes		Yes	
Class Covariates	Yes		Yes	
School FE	Yes		Yes	

Note: This table reports the result of a negative binomial regression separately for low-, middle- and high-ability pupils and separately for High Schools and Vocational School including school fixed effects. Dependent variable: points in test; Covariates: last midterm grade, gender, number of books at home, academic year (grade 5 or 6), teachers' working experience (in years), teacher's gender, day differences between tests and the proportion of German speaking pupils within the class. High-ability pupils refers to those with a midterm grade of 1 or 2; middle-ability pupils have a midterm grade of 3 and low-ability pupils are those with a midterm grade of 4, 5 or 6. The groups are of approximately equal size. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 36 in High Schools and 53 in Vocational Schools. * p<0.10, ** p<0.05, *** p<0.01.

Table 2.13: Treatment Effects by Pupils' Gender and Midterm Grade

	Low Performers				**Middle Performers**				**High Performers**			
Panel A: Regression	*Vocational School*		*High School*		*Vocational School*		*High School*		*Vocational School*		*High School*	
Treatments												
Choice	-0.322	[1.538]	-0.856	[2.926]	0.752	[2.170]	0.783	[2.195]	2.156	[1.956]	-0.392	[1.311]
Medal	-2.444	[1.599]	-3.946	[3.265]	-3.469*	[1.940]	-2.275	[1.836]	2.558*	[1.446]	0.107	[1.502]
Letter	-0.570	[1.915]	-4.672	[2.955]	-0.544	[1.892]	-4.912**	[2.286]	2.590	[1.853]	-2.877	[1.922]
Female	-3.217**	[1.357]	-0.787	[3.736]	-1.178	[1.603]	0.420	[1.247]	0.639	[2.746]	-1.940	[1.570]
Choice × Female	1.074	[2.222]	-3.872	[3.806]	-1.468	[2.436]	0.020	[3.126]	-1.362	[3.274]	2.275	[2.200]
Medal × Female	6.542***	[2.265]	-0.649	[4.798]	-1.584	[2.626]	-2.193	[2.893]	-3.693	[3.019]	-1.118	[2.661]
Letter × Female	2.120	[2.806]	-1.536	[4.362]	-1.786	[2.247]	2.679	[2.551]	0.817	[4.321]	1.589	[1.962]
Controls												
Pupil Covariates	Yes		Yes		Yes		Yes		Yes		Yes	
Class Covariates	Yes		Yes		Yes		Yes		Yes		Yes	
School FE	Yes		Yes		Yes		Yes		Yes		Yes	
Panel B: Contrast	*Treatment vs. No Treatment for Females*											
Choice	0.751	[1.599]	-4.727*	[2.533]	-0.716	[2.396]	0.803	[2.559]	0.794	[2.169]	1.883	[1.624]
Medal	4.098**	[1.847]	-4.595	[2.927]	-5.053**	[2.250]	-4.468*	[2.521]	-1.134	[2.400]	-1.010	[2.256]
Letter	1.550	[2.022]	-6.208*	[3.324]	-2.330	[2.340]	-2.234	[2.359]	3.407	[3.018]	-1.288	[2.288]
N	408		144		428		288		362		437	

Note: Panel A reports the result of a negative binomial regression separately for low-, middle- and high-ability boys and separately for High Schools and Vocational Schools and also reports interaction effects with gender including school fixed effects. Panel B reports treatment effect for girls. Dependent variable: points in test; Covariates: last midterm grade, number of books at home, academic year (grade 5 or 6), teachers' working experience (in years), teacher's gender, day differences between tests and the proportion of German speaking pupils within the class. Female: 0 = boys; 1 = girls. High-ability pupils refers to those with a midterm grade of 1 or 2; middle-ability pupils have a midterm grade of 3 and low-ability pupils are those with a midterm grade of 4, 5 or 6. The groups are of approximately equal size. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 36 in High Schools and 53 in Vocational Schools. * p<0.10, ** p<0.05, *** p<0.01.

Test preparation

Table 2.14: Test Preparation by School Type

Panel A: Regression	*Vocational School*		*High School*	
Treatments				
Choice	0.208**	[0.082]	0.136***	[0.051]
Medal	0.186*	[0.102]	0.012	[0.045]
Letter	-0.095	[0.092]	0.075**	[0.038]
Grade 6	0.009	[0.127]	-0.335***	[0.052]
Choice × Grade 6	-0.112	[0.140]	0.037	[0.082]
Medal × Grade 6	-0.147	[0.158]	0.044	[0.177]
Letter × Grade 6	0.159	[0.160]	0.090	[0.058]
Controls				
Female pupil	0.130***	[0.028]	0.080**	[0.038]
Midterm grade	0.007	[0.005]	0.027***	[0.008]
Like Maths	0.040***	[0.013]	0.052**	[0.021]
School FE	Yes		Yes	
N	1189		866	

Panel B: Contrasts *Treatment vs. No Treatment in Year 6*				
Choice	0.096	[0.101]	0.173***	[0.059]
Medal	0.039	[0.117]	0.057	[0.172]
Letter	0.064	[0.115]	0.166***	[0.042]

Note: Panel A reports results of logistic regression (marginal effects) for pupils in grade 5 and the interaction terms for treatment and school level separately for Vocational Schools and High Schools including school fixed effects. Panel B reports the logistic treatment effects for pupils in grade 6. Grade 6: 0=pupils in grade 5, 1=pupils in grade 6. Dependent variable: prepared for test (*Did you prepare for the test?* 0=No, 1=Yes). Covariates: last midterm grade, math curiosity (measured on 1 to 5 scale). Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. * p<0.10, ** p<0.05, *** p<0.01.

Effects of covariates

Table 2.15: Ordinary Least Squares Regression of Covariates on Test Performance

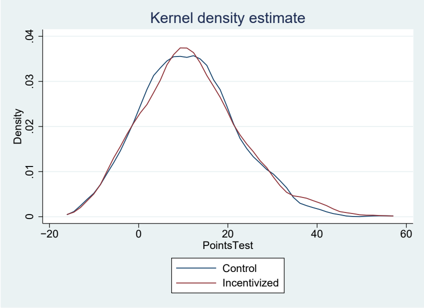
	Overall		*Male*		*Female*		*High Performers*		*Low Performers*	
Panel A: Vocational Schools										
BooksHome (11-25)	1.200	[0.848]	2.461**	[1.020]	-0.100	[1.231]	2.536	[1.574]	-0.435	[1.805]
BooksHome (26-100)	1.673*	[0.939]	2.621**	[1.144]	0.760	[1.288]	3.718*	[1.902]	-0.061	[1.854]
BooksHome (101-200)	2.734***	[1.028]	2.062	[1.470]	3.331**	[1.568]	4.305*	[2.322]	0.419	[1.982]
BooksHome (201-500)	3.375**	[1.465]	4.389**	[2.136]	2.500	[2.347]	4.195*	[2.455]	2.391	[2.626]
BooksHome (over 500)	2.747*	[1.503]	3.256	[2.271]	2.515	[2.293]	1.934	[3.268]	0.904	[3.330]
BooksHome (Not Reported)	3.310***	[1.254]	3.882**	[2.038]	3.066**	[1.340]	1.982	[3.686]	1.831	[2.460]
GradeMidTerm	-1.288***	[0.098]	-1.295***	[0.141]	-1.248***	[0.141]	-2.610***	[0.621]	-0.697**	[0.346]
TeacherExperience	0.090**	[0.037]	0.061	[0.059]	0.150***	[0.048]	-0.006	[0.039]	0.181***	[0.044]
TeacherFemale	-1.099	[0.978]	0.429	[0.945]	-2.909**	[1.309]	1.812	[1.480]	-2.449**	[1.134]
N	1198		665		533		362		408	
Panel B: High Schools										
BooksHome (11-25)	3.071*	[1.768]	4.552**	[1.982]	2.834	[3.509]	4.134	[3.245]	2.165	[2.165]
BooksHome (26-100)	3.953**	[1.627]	3.695**	[1.848]	5.846	[3.745]	3.282	[2.710]	4.976*	[2.919]
BooksHome (101-200)	2.490	[1.674]	3.056	[2.047]	3.938	[4.014]	2.145	[2.806]	5.006	[3.305]
BooksHome (201-500)	5.065***	[1.614]	4.990***	[1.788]	7.099*	[4.049]	4.970*	[2.537]	6.797*	[3.622]
BooksHome (over 500)	5.095***	[1.720]	5.254**	[2.079]	7.680*	[4.131]	5.173*	[2.831]	6.027*	[3.201]
BooksHome (Not Reported)	2.936	[1.828]	3.473*	[1.954]	4.943	[4.772]	2.700	[]2.499	1.238	[3.593]
GradeMidTerm	-1.441***	[0.129]	-1.581***	[0.163]	-1.214***	[0.179]	-2.646***	[0.319]	-0.741*	[0.417]
TeacherExperience	0.058*	[0.032]	0.077**	[0.034]	0.019	[0.048]	0.085	[0.063]	0.165**	[0.079]
TeacherFemale	-1.912**	[0.809]	-3.109***	[0.800]	-0.387	[1.148]	-1.808	[1.139]	-3.854**	[1.853]
N	869		504		365		437		144	

Note: Panel A reports results of negative binomial regression (marginal effects) of the covariates for pupils in Vocational School including school fixed effects. Panel B results of negative binomial regression (marginal effects) of the covariates for pupils in High School including school fixed effects. Dependent variable of the initial regression is points in test. Baseline for books at home is the category 0-10. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 53 in Vocational Schools and 36 in High Schools. * p<0.10, ** p<0.05, *** p<0.01.

Kernel density plots by Treatment

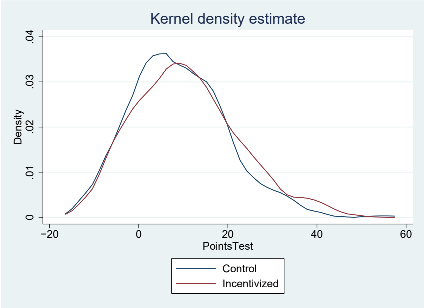
Figure 2.2: Kernel Density Estimation Control vs. Incentivized

All School Types

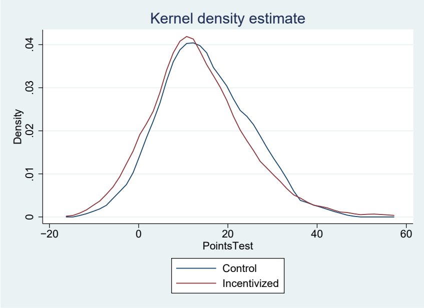


Note: This Figure presents kernel density estimates for the test performance for incentivized and not incentivized pupils pooled over school types.

Vocational School (a)



High School (b)



Note: Figure (b) presents kernel density estimates for the test performance for incentivized and not incentivized pupils in Vocational Schools. Figure (c) presents kernel density estimates for the test performance for incentivized and not incentivized pupils in High Schools.

German School System

In the German school system, children are segregated into high and low performers at an early age. Elementary school in Germany runs from grade one at the age of 6 to grade four at the age of 9 or 10. School-aged children must attend the school in their school district.[36] With the semester report in grade four, parents receive a transition recommendation to which school type to send their child. This recommendation is given by the elementary school and is based on talent and performance (i.e., grades), social skills and social behavior and motivation and learning virtues (Anders et al. 2010). However, parents in NRW can decide to which type of secondary school they want to send their children, regardless of the recommendation. Nevertheless, depending on their capacity, secondary schools can decline applications.[37]

The German secondary school system consists mainly of four school types (approximate US equivalents in parenthesis): *Hauptschule* (Secondary General School), *Realschule* (Middle School), *Gesamtschule* (Comprehensive School) and *Gymnasium* (High School). In the following, we use the US equivalents. The average class size consists of 21–28 pupils and a typical week for fifth and sixth graders consists of approximately 37 school hours.[38] Typically, pupils remain in the same class from grade 5 until grade 10, at which time they turn into a course system. Therefore, classes are closed units in which most of the social interaction in pupils' school life takes place. The German grade system ranges from 1 to 6, in which 1 is the highest possible grade and 5 is the threshold for failing—the US equivalents are A+ to F.

The *Secondary General School* (grades five to nine or ten) provides pupils with a basic general education that prepares them, in particular, for a vocational job and finishes with a *Hauptschulabschluss* after grade nine or ten. Depending on performance, pupils can qualify to attend the advanced level of High School.

The *Middle School* (grades five to ten) encourages practical skills as well as interest in theoretical context. Pupils acquire an advanced education and career guidance skills. Furthermore, in grade six, pupils learn a second foreign language. After completion of the tenth grade—and depending on past performance and interest—pupils can change to a vocational training course or attend the advanced level of the High School if his/her grades are good enough. The minimum grade for continuing to High School is 3 on average in all subjects.[39]

[36] In 2008, the forced allocation of pupils to the elementary school in their specific district, determined by address, was abolished in the federal state of North Rhine-Westphalia—where we conducted our experiment. This means that parents of the cohorts in our study were free to decide to which elementary school they sent their children.

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

[38] This information is taken from the Ministry of Education and Further Education of NRW. For further information see `http://www.schulministerium.nrw.de/docs/bp/Ministerium/Service/Schulstatistik/Amtliche-Schuldaten/StatTelegramm2012.pdf`.

[39] A sufficient performance (grade 4) in a major subject—German, Mathematics, English—can be compensated by a good performance (grade 2) in another major subject. A maximum of three sufficient performances in a minor subject or two sufficient and one poor performance (grade 5) can be compensated for by an equal number of good performance in other minor subjects.

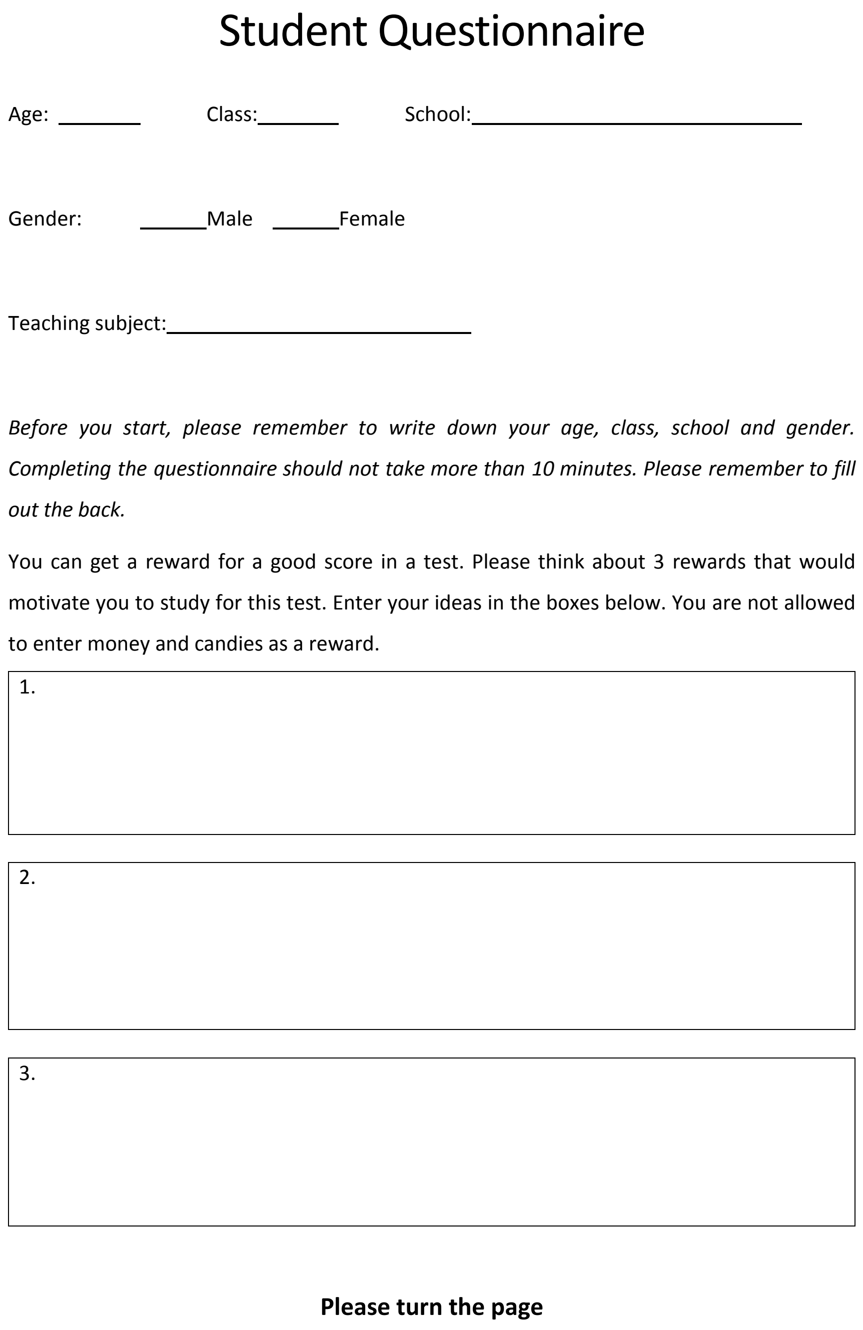
Children attending the *Comprehensive School* (grades five to ten or twelve) have a longer period of common learning. Classes consist of children of all skill levels and career decisions are left open as long as possible. The majority of *Comprehensive Schools* are all-day schools in which all degrees of secondary education can be achieved that are awarded at Secondary General School, Middle School and High School. As with the *Middle School*, pupils can qualify for the advanced level of the High School and obtain the Abitur (A-Level).

The *High School* (grades five to twelve) is the most academic school type. The final examination—the Abitur—entitles students to apply to University.[40] The aim of the High School is to give an in-depth general education, which is necessary for both higher education and for vocational training. The lessons should guide the analysis of complex problems and lead to abstraction, analytical and critical thinking capabilities.

[40] All other education degrees from Secondary General School and Middle School can also be acquired at the High School.

Survey

Figure 2.3: Survey Front and Back



Student Questionnaire

Age: ________ Class:________ School:________________________________

Gender: ______Male ______Female

Teaching subject:____________________________

Before you start, please remember to write down your age, class, school and gender. Completing the questionnaire should not take more than 10 minutes. Please remember to fill out the back.

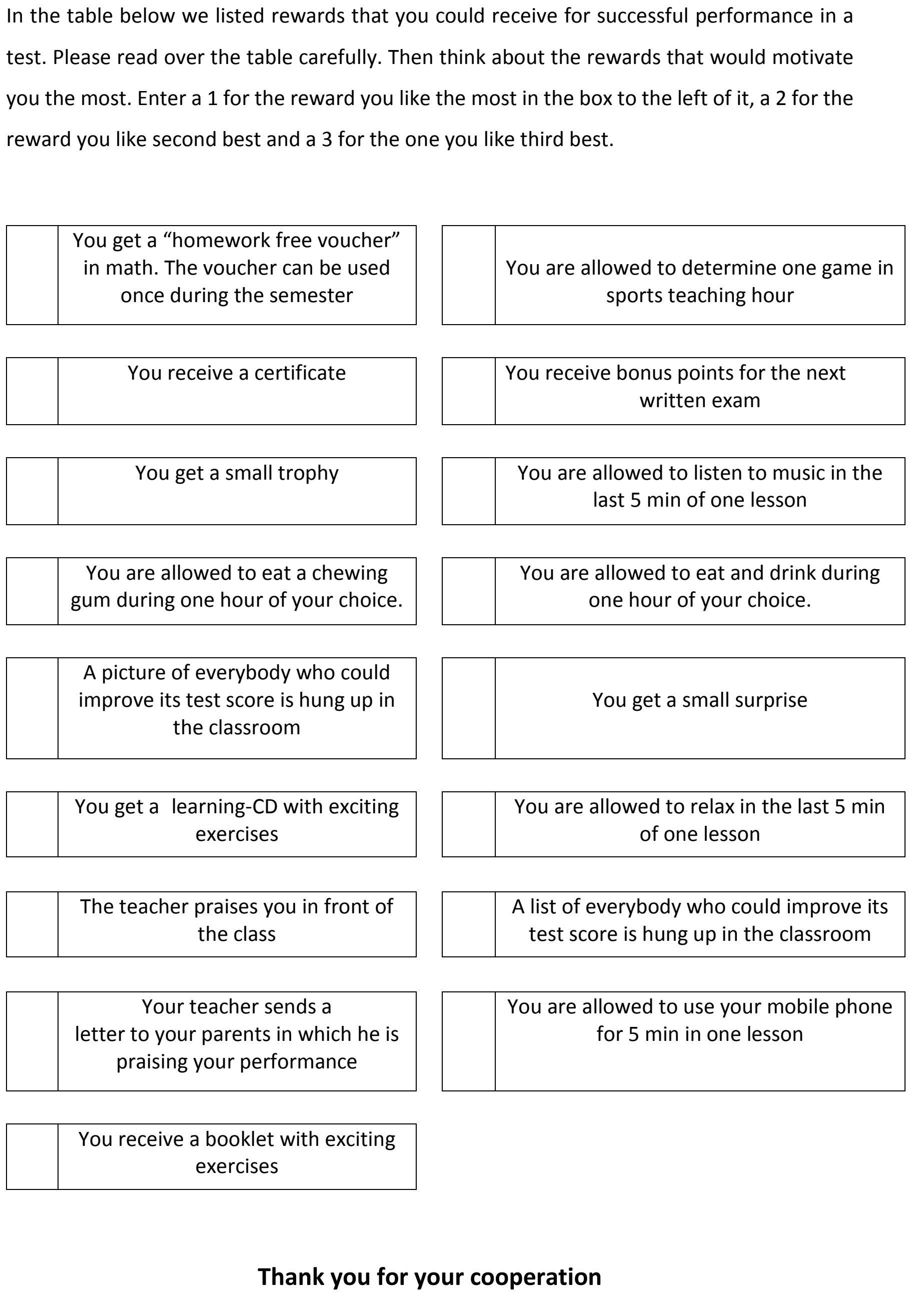
You can get a reward for a good score in a test. Please think about 3 rewards that would motivate you to study for this test. Enter your ideas in the boxes below. You are not allowed to enter money and candies as a reward.

1.

2.

3.

Please turn the page



In the table below we listed rewards that you could receive for successful performance in a test. Please read over the table carefully. Then think about the rewards that would motivate you the most. Enter a 1 for the reward you like the most in the box to the left of it, a 2 for the reward you like second best and a 3 for the one you like third best.

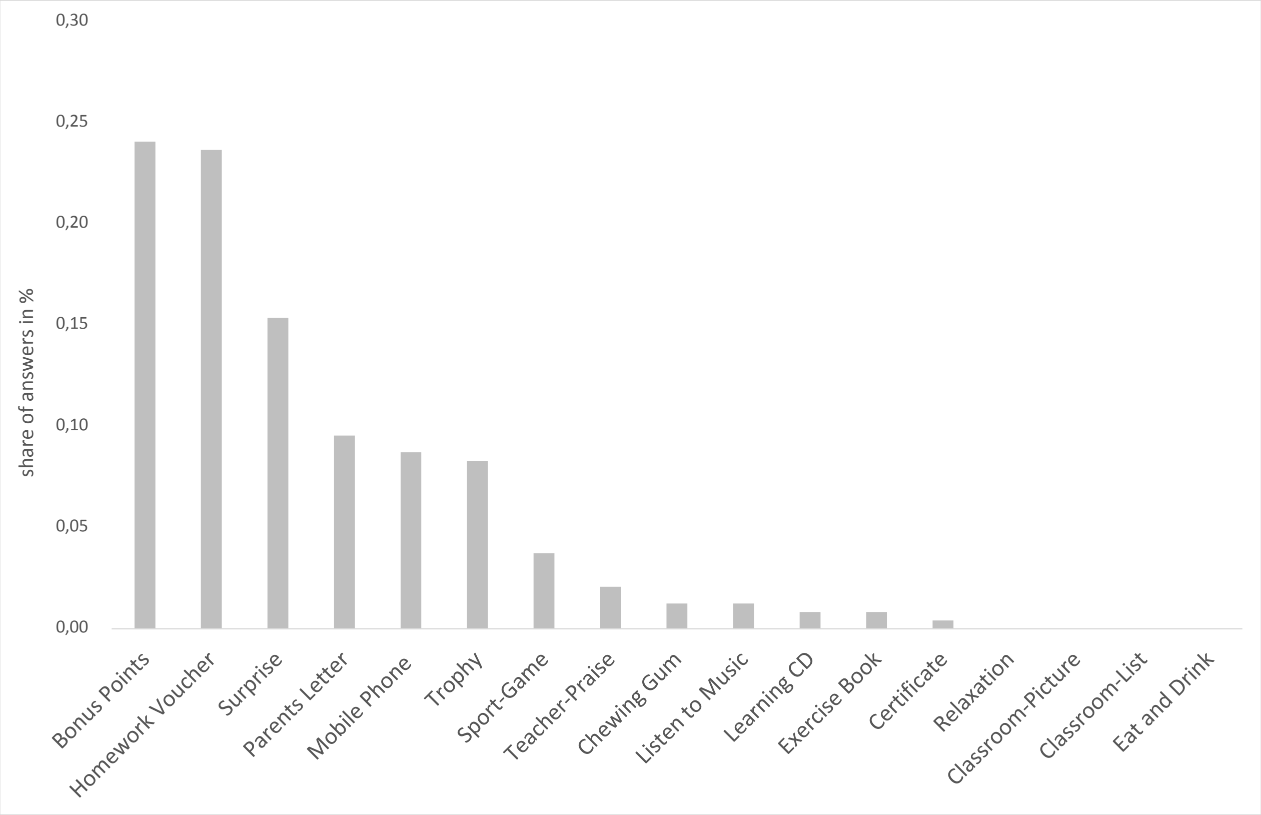
	You get a “homework free voucher” in math. The voucher can be used once during the semester		You are allowed to determine one game in sports teaching hour
	You receive a certificate		You receive bonus points for the next written exam
	You get a small trophy		You are allowed to listen to music in the last 5 min of one lesson
	You are allowed to eat a chewing gum during one hour of your choice.		You are allowed to eat and drink during one hour of your choice.
	A picture of everybody who could improve its test score is hung up in the classroom		You get a small surprise
	You get a learning-CD with exciting exercises		You are allowed to relax in the last 5 min of one lesson
	The teacher praises you in front of the class		A list of everybody who could improve its test score is hung up in the classroom
	Your teacher sends a letter to your parents in which he is praising your performance		You are allowed to use your mobile phone for 5 min in one lesson
	You receive a booklet with exciting exercises		

Thank you for your cooperation

Table 2.16: Predetermined Incentives given in Survey

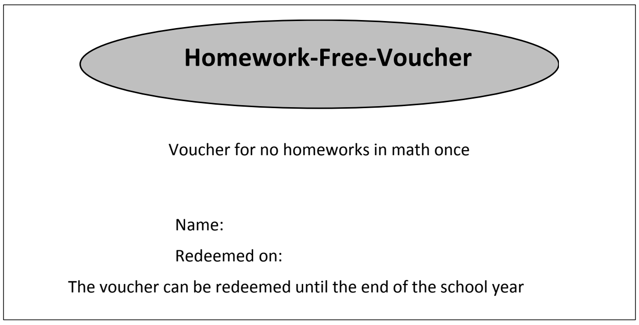
Work Avoidance	Mastery Goal	Social Recocnition		Consumption	Curiosity
		Public	Private		
Homework voucher: No homework in math. Voucher can be used once until the end of semester.	*Exercise Book*: Receiving a booklet with mathematical exercises	*Teacher-Praise*: Being praised in front of the class	*Parents-Letter*: Teacher sends letter to parents, praising the pupil's performance	*Chewing Gum:* Being allowed to eat a chewing gum in one lesson	*Surprise*: Getting a small surprise reward
Relaxation: Relaxing 5 minutes of one lesson	*Learning CD*: Receiving a CD with mathematical games	*Classroom-Picture*: Picture of pupil who could improve its test score is hung up in classroom	*Trophy*: Getting a small Medal	*Listen to Music*: Being allowed to listen to music for 5 min	
Bonus points: Receiving extra points for next written exam		*Classroom-List*: List of all who could improve their test score is hung up in classroom	*Certificate:* Receiving a certificate stating that test score could be improved	*Eat and Drink:* Being allowed to eat and drink during class	
				Mobile Phone: Being allowed to play 5 min with mobile in one lesson	
				Sport-Game: Determing one game in sports teaching hour	

Figure 2.4: Survey Answers



Facsimile of Incentives

Figure 2.5: Example of No-Homework-Voucher



Homework-Free-Voucher

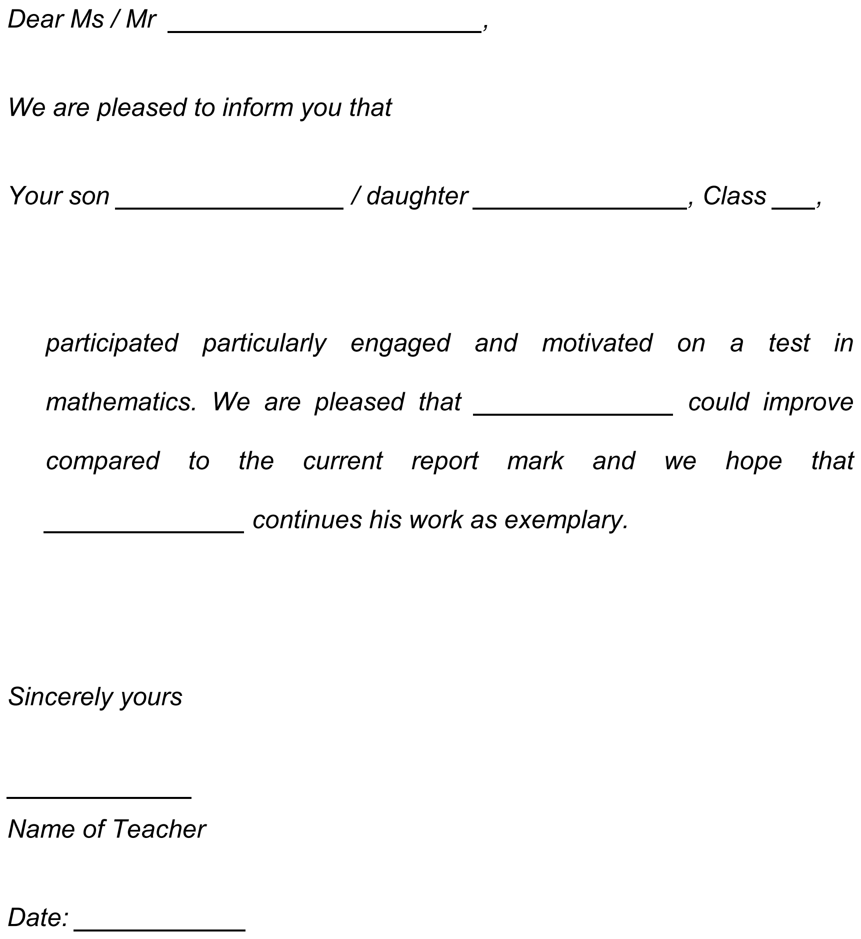
Voucher for no homeworks in math once

Name:

Redeemed on:

The voucher can be redeemed until the end of the school year

Figure 2.6: Example of Parent-Letter



Dear Ms / Mr ______________________,

We are pleased to inform you that

Your son ________________ / daughter _______________, Class ___,

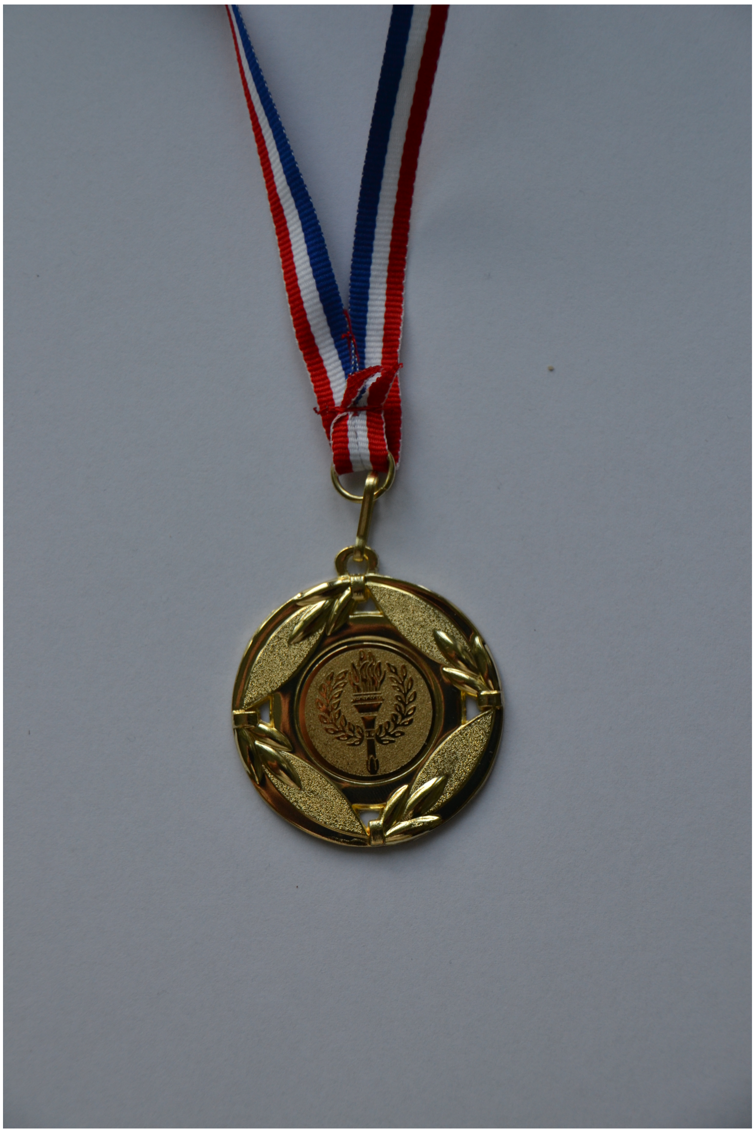
participated particularly engaged and motivated on a test in mathematics. We are pleased that ______________ could improve compared to the current report mark and we hope that ______________ continues his work as exemplary.

Sincerely yours

____
Name of Teacher

Date: ____________

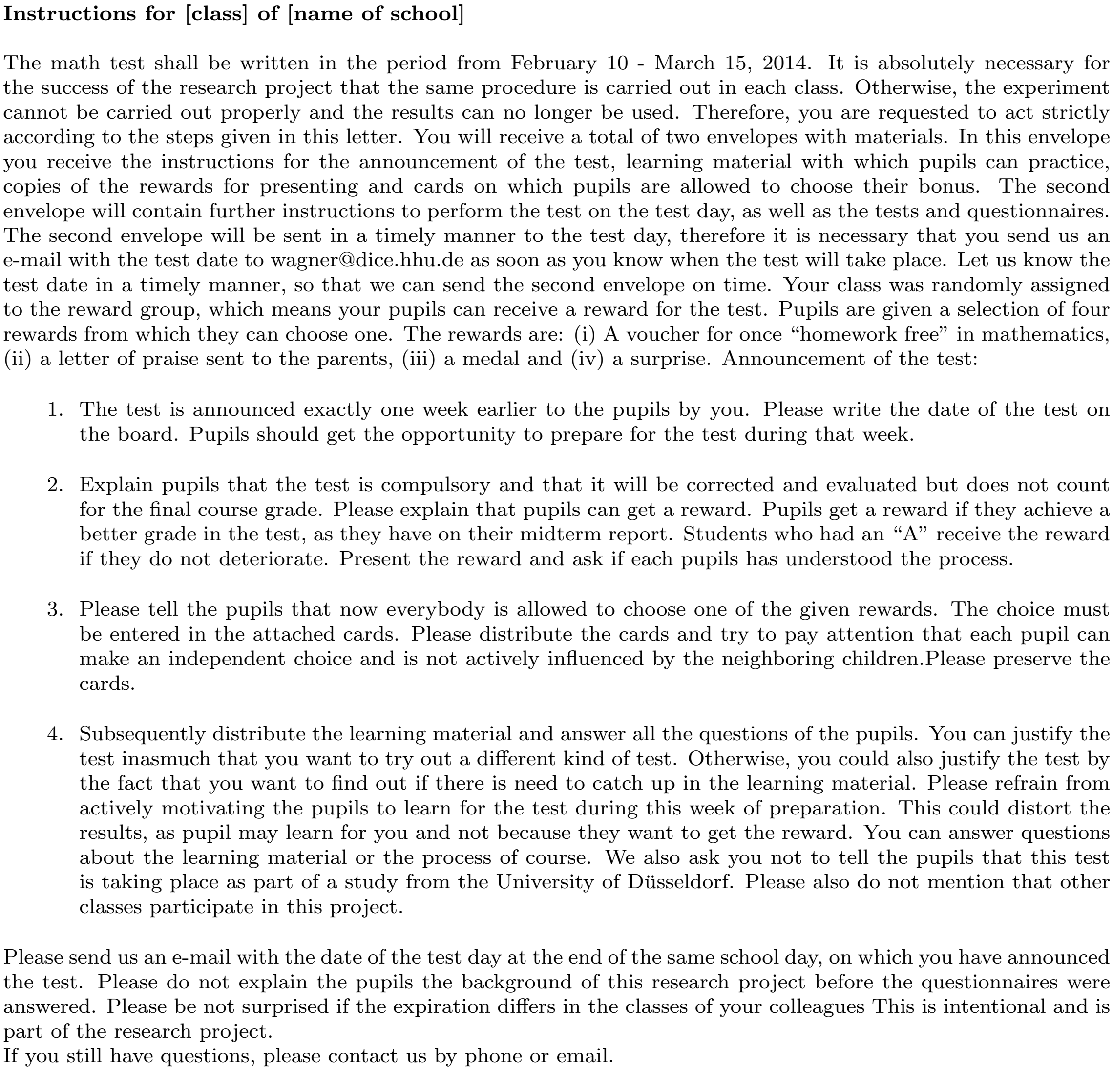
Figure 2.7: Picture of Medal



Teacher Instructions

Figure 2.8 shows instructions for teachers in the Choice Treatment. Instructions for the Control and Fixed Treatments are similar but the text is slightly adjusted to the respective incentives. Furthermore, point 3 was not included in the instructions of the Control and Fixed Treatments.

Figure 2.8: Instructions for Teachers - First Mailing [Translated from German]



Instructions for [class] of [name of school]

The math test shall be written in the period from February 10 - March 15, 2014. It is absolutely necessary for the success of the research project that the same procedure is carried out in each class. Otherwise, the experiment cannot be carried out properly and the results can no longer be used. Therefore, you are requested to act strictly according to the steps given in this letter. You will receive a total of two envelopes with materials. In this envelope you receive the instructions for the announcement of the test, learning material with which pupils can practice, copies of the rewards for presenting and cards on which pupils are allowed to choose their bonus. The second envelope will contain further instructions to perform the test on the test day, as well as the tests and questionnaires. The second envelope will be sent in a timely manner to the test day, therefore it is necessary that you send us an e-mail with the test date to wagner@dice.hhu.de as soon as you know when the test will take place. Let us know the test date in a timely manner, so that we can send the second envelope on time. Your class was randomly assigned to the reward group, which means your pupils can receive a reward for the test. Pupils are given a selection of four rewards from which they can choose one. The rewards are: (i) A voucher for once "homework free" in mathematics, (ii) a letter of praise sent to the parents, (iii) a medal and (iv) a surprise. Announcement of the test:

1. The test is announced exactly one week earlier to the pupils by you. Please write the date of the test on the board. Pupils should get the opportunity to prepare for the test during that week.

2. Explain pupils that the test is compulsory and that it will be corrected and evaluated but does not count for the final course grade. Please explain that pupils can get a reward. Pupils get a reward if they achieve a better grade in the test, as they have on their midterm report. Students who had an "A" receive the reward if they do not deteriorate. Present the reward and ask if each pupils has understood the process.

3. Please tell the pupils that now everybody is allowed to choose one of the given rewards. The choice must be entered in the attached cards. Please distribute the cards and try to pay attention that each pupil can make an independent choice and is not actively influenced by the neighboring children.Please preserve the cards.

4. Subsequently distribute the learning material and answer all the questions of the pupils. You can justify the test inasmuch that you want to try out a different kind of test. Otherwise, you could also justify the test by the fact that you want to find out if there is need to catch up in the learning material. Please refrain from actively motivating the pupils to learn for the test during this week of preparation. This could distort the results, as pupil may learn for you and not because they want to get the reward. You can answer questions about the learning material or the process of course. We also ask you not to tell the pupils that this test is taking place as part of a study from the University of Düsseldorf. Please also do not mention that other classes participate in this project.

Please send us an e-mail with the date of the test day at the end of the same school day, on which you have announced the test. Please do not explain the pupils the background of this research project before the questionnaires were answered. Please be not surprised if the expiration differs in the classes of your colleagues This is intentional and is part of the research project.
If you still have questions, please contact us by phone or email.

Figure 2.9 shows instructions for teachers on the testing day in the Choice and Fixed Treatments. Instructions for the Control Treatment were the same except that point 3 was not included.

Figure 2.9: Instructions for Teachers - Second Mailing [Translated from German]

Instructions for [class] of [name of school]

With this envelope you receive the tests, questionnaires, a list to enter the midterm grades and a statement of privacy. Please read the instructions carefully and carry out the test in the given steps.

Implementation of the tests: Test-duration 30 minutes

1. Please let the pupils - similar to exams - set the tables a little bit apart. Additionally let them put up a privacy screen between each other. Remind the pupils that all questions have to be answered independently and that each attempt to copy from the test 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 14 tasks that must be solved within 30 minutes. For each task, there are 4 wrong and 1 correct answers. There are tasks that are worth 3 points for each correct answer, and others that are worth 4 or 5 points. If an incorrect answer is written, 1 point is deducted. If no answer is given, you receive 0 points. Calculators are not allowed, but “scratch paper” for sketches and small calculations are allowed, of course!”

3. Please, remind the pupils of the rewards one more time and present them to the class. Mention also that it will take no longer than one week until the tests are evaluated and pupils receive their rewards. It is important for the motivation of the pupils that they know that the rewards will be distributed in a timely manner.

4. Please tell the pupils that they should not write their names on the test. For privacy reasons, each test receives a “Test-ID number”.

5. Now the test starts and lasts for 30 minutes in total.

6. 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 separate sheet of paper. For this, you could also use a class list. This sheet serves as an “encryption key” that 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.

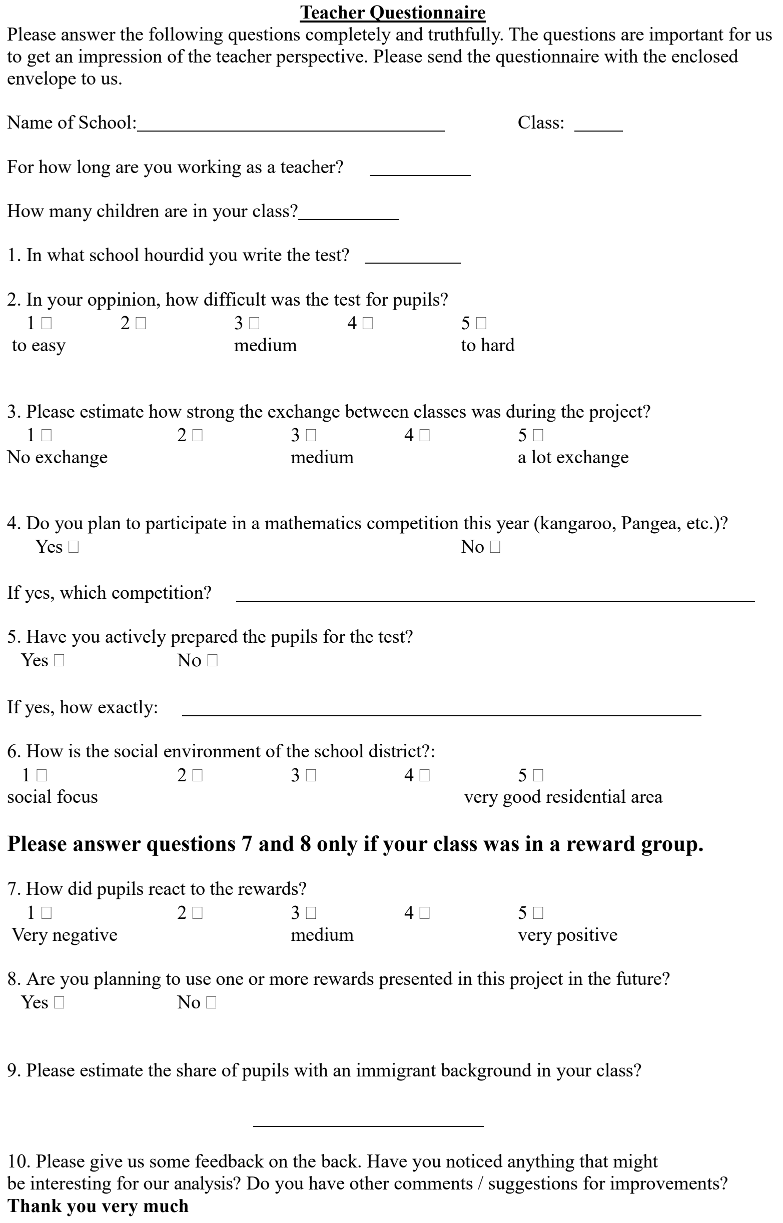
7. After the test, the questionnaires are answered. These have already been attached to the test. Again, the questionnaire has to be answered independently and quietly.

8. Please collect the preparation sheets after each pupil has responded to the questionnaire and write the corresponding ID number on it. Pupils who do not have their exercise sheets with them, shall hand them in during the next week. Based on the exercise sheets we want to see if pupils have worked on the tasks.

Please send the tests, questionnaires, preparation sheets and the list with the midterm grade back to us with the enclosed envelope on the same day. The tests are then corrected immediately thereafter and sent back to you with the rewards. Please fill in the midterm grades and chosen incentives in the list we have send to you. The Test-ID numbers serve as an encryption key. Example: “Andrea Albers”, has the Test-ID number 12 and has chosen the medal, then please write down under the number 12, the midterm grade plus tendency of Andrea Albers and that she has chosen the medal. By this method, we can meet the requirements of privacy policy since it cannot be identified retrospectively which grade belongs to which pupil. In addition, all materials that 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. This will take some time, we expect a full analysis in June / July 2014. Thank you very much

Teacher and Pupil Questionnaire

Figure 2.10: Teacher Questionnaire [Translated from German]



Teacher Questionnaire

Please answer the following questions completely and truthfully. The questions are important for us to get an impression of the teacher perspective. Please send the questionnaire with the enclosed envelope to us.

Name of School:________________________________ Class: _____

For how long are you working as a teacher? ___________

How many children are in your class?___________

1. In what school hourdid you write the test? ___________

2. In your oppinion, how difficult was the test for pupils?
1 □ 2 □ 3 □ 4 □ 5 □
to easy medium to hard

3. Please estimate how strong the exchange between classes was during the project?
1 □ 2 □ 3 □ 4 □ 5 □
No exchange medium a lot exchange

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

If yes, which competition? ____

5. Have you actively prepared the pupils for the test?
Yes □ No □

If yes, how exactly: ____

6. How is the social environment of the school district?:
1 □ 2 □ 3 □ 4 □ 5 □
social focus very good residential area

Please answer questions 7 and 8 only if your class was in a reward group.

7. How did pupils react to the rewards?
1 □ 2 □ 3 □ 4 □ 5 □
Very negative medium very positive

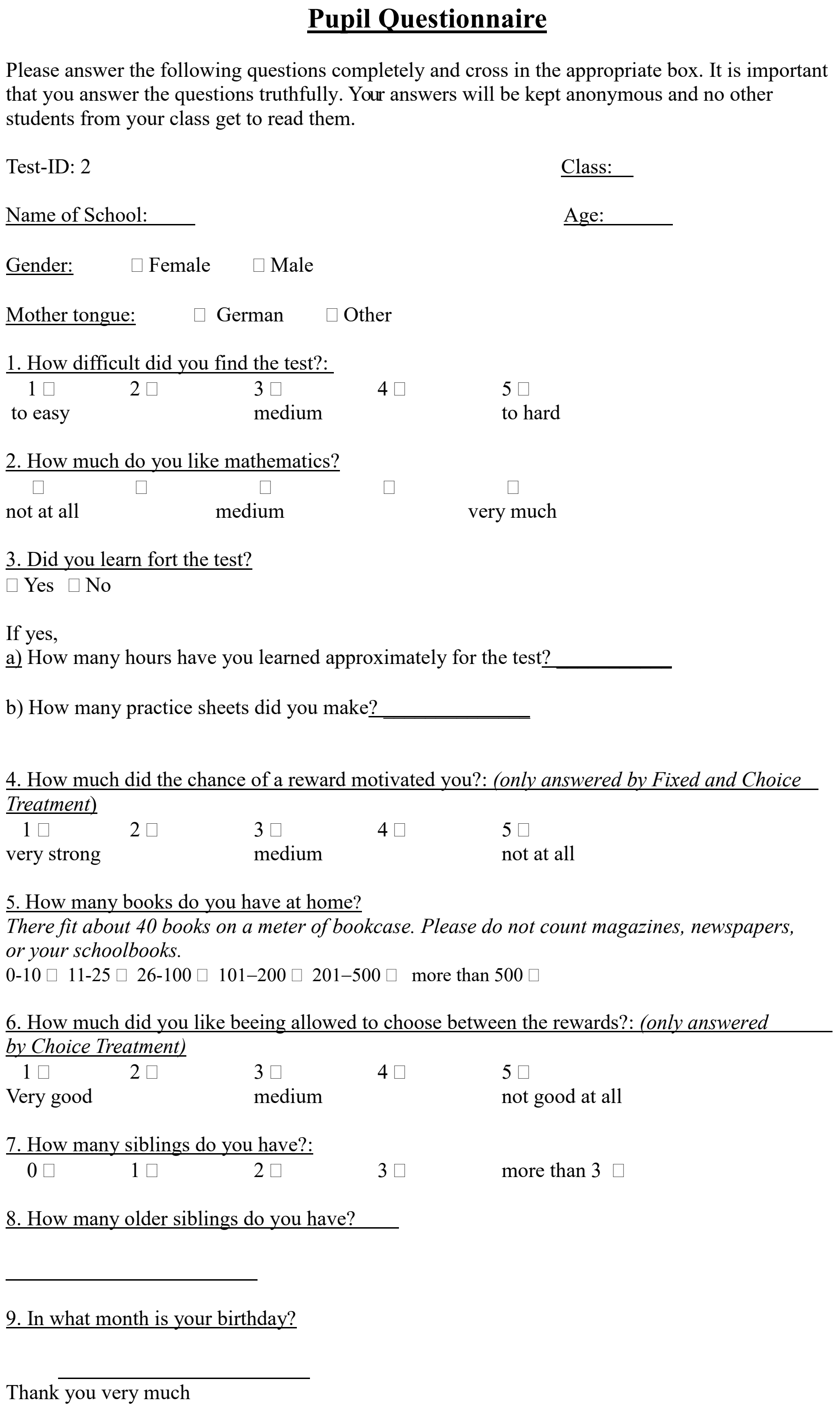
8. Are you planning to use one or more rewards presented in this project in the future?
Yes □ No □

9. Please estimate the share of pupils with an immigrant background in your class?

____

10. Please give us some feedback on the back. Have you noticed anything that might
be interesting for our analysis? Do you have other comments / suggestions for improvements?
Thank you very much

Figure 2.11: Pupil Questionnaire [Translated from German]



Pupil Questionnaire

Please answer the following questions completely and cross in the appropriate box. It is important that you answer the questions truthfully. Your answers will be kept anonymous and no other students from your class get to read them.

Test-ID: 2 Class: __

Name of School: _____ Age: _______

Gender: □ Female □ Male

Mother tongue: □ German □ Other

1. How difficult did you find the test?:

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

2. How much do you like mathematics?

□ □ □ □ □
not at all medium very much

3. Did you learn fort the test?

□ Yes □ No

If yes,
a) How many hours have you learned approximately for the test?___________

b) How many practice sheets did you make?______________

4. How much did the chance of a reward motivated you?: *(only answered by Fixed and Choice Treatment)*

1 □ 2 □ 3 □ 4 □ 5 □
very strong medium not at all

5. How many books do you have at home?
There fit about 40 books on a meter of bookcase. Please do not count magazines, newspapers, or your schoolbooks.
0-10 □ 11-25 □ 26-100 □ 101–200 □ 201–500 □ more than 500 □

6. How much did you like beeing allowed to choose between the rewards?: *(only answered by Choice Treatment)*

1 □ 2 □ 3 □ 4 □ 5 □
Very good medium not good at all

7. How many siblings do you have?:

0 □ 1 □ 2 □ 3 □ more than 3 □

8. How many older siblings do you have? ____

____

9. In what month is your birthday?

____

Thank you very much

Chapter 3

Seeking Risk or Answering Smart? Framing in Elementary Schools

3.1 Introduction

Effort is an important prerequisite to achieve externally imposed goals. Managers may set a goal for productivity in the workplace, doctors advise their patient 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] Moreover, inducing loss aversion to change peoples' behavior tends to be effective and hence the framing of extrinsic rewards as a loss has been increasingly applied to some field settings in recent years (Hong et al. 2015; Armantier and Boly 2015; List and Samek 2015; Fryer et al. 2012; Hossain and List 2012). These studies demonstrate that the provision of effort is sensitive to incentives framing. However, it is important to know for whom loss framing works and to understand the underlying mechanisms of effort provision if outcomes depend on multiple inputs i.e. the quality and quantity of decisions.

An ideal setting to test the impact of framing effects on the quality and quantity of decisions is within the educational sector 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.[2] It also allows to analyze 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 although there are large returns to education (Hanushek et al. 2015; Card and Krueger 1992; Card 1999).[3] To test framing effects is therefore promising as it represents a potential cost-effective and easy to implement method to motivate pupils. 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.[4] Therefore, enhancing pupils' positive attitude towards school (i) might be more effective in younger ages due to complementarities of skill formation

[1] Springer et al. (2015); Jalava et al. (2015); Levitt et al. (2016) and Chapter 2 analyze the effectiveness of non-monetary incentives in educational settings.

[2] 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).

[3] See Lavecchia et al. (2016) and Koch et al. (2015) for an overview on behavioral economics of education.

[4] A more detailed description of the German tracking system is given in Chapter 2.

at different stages of the education production function (Cunha and Heckman 2007) and (ii) might influence the tracking decision and thus pupils' future income.[5]

Pupils in elementary schools represent the general population as they are not yet tracked by ability and, based on their midterm grades, they can be differentiated into high-, middle- and low-performers.[6] While high-performers are likely to be allocated to the academic track and low-performers to the lower track (preparing for blue color occupations), middle-ability pupils might the most at 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.[7] 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 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 ten item multiple-choice test and compares pupils' answering behavior under three different frames: (1) gain frame, (2) loss frame and (3) gain frame with negative endowment. Moreover, a special focus is on analyzing the effectiveness of framing effects for different ability levels (high- and low-performing pupils). To the best of my knowledge this has not been studied previously and it represents a major contribution of this paper. Furthermore, the multiple-choice testing format allows to analyze the impact of framing effects on pupils' risk-seeking behavior and level of accuracy.[8]

The experiment was conducted in 20 elementary schools in Germany among 1.377 pupils of grades 3 and 4. The setting of elementary schools allows to analyze framing effects for 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 *Loss Treatment* and the *Negative 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

[5] 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.

[6] Pupils usually attend the elementary school which is in their close neighborhood.

[7] 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 of the experiment.

[8] 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%.

skipping an answer and 0 points for an incorrect answer.[9] These two treatments differ with respect to pupils' initial endowment—either 0 points or -20 points. Hence, pupils could earn between 0 to 40 points in the Control Group and -20 to +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".[10] 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 answer and 0 points for a correct answer.

On average, 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.[11] This can be interpreted as an increased risk-seeking behavior of pupils in the Loss Treatment and an increase in accuracy of pupils in the Negative Treatment. Moreover, I find heterogeneous framing effects for pupils of different ability levels. While high-ability pupils increase the number of correct answers as well as 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 to separate the effort effect from the learning effect. 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 3.3 and Section 3.4 derives hypotheses of potential treatment effects. The data and descriptive statistics are reported in Section 3.5. Section 3.6 presents the results which are discussed in Section 3.7. Section 3.8 summarizes and concludes.

[9] 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.

[10] This information was informally given by teachers.

[11] Overall, the coefficient for the number of *total points* in the test is positive but statistical insignificant for both treatments.

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

3.2 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 researcher as—compared to financial incentives—this 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 kids 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 in Chapter 2 we have tested 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 a "traditional" criterion-based grading (pupils get grade on a A-F scale according to predetermined thresholds) and a rank-based grading. In the latter, only the top three performers of a class received an A. The authors find that rank-based grading increases performance of boys and girls and that rank-based grading also tends to crowd out 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-stakes testing environment among university students. The authors hypothesize that grading on a curve induces male students to increase their performance compared to an absolute grading. They find weak support for this hypothesis and 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 depends on the peer group composition, i.e. 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 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 in educational settings. 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.[15] Apostolova-Mihaylova et al. (2015) test whether framing grades of university students as a

[13] See also Bradler et al. (2016); Bradler and Neckermann (2016); Ashraf et al. (2014); Neckermann et al. (2014); Kube et al. (2012); Goerg and Kube (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).

[15] The size of gains was equivalent to increasing teacher quality by more than one standard deviation.

loss or as a gain effects 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.[16] After each completed exam or assignment, the students' grades were updated, so that students had the opportunity to follow their increasing or decreasing grades. The authors find no overall effect of loss framing on the final course grade but they find heterogeneous gender effects. The final course grade of male students increased while female students got lower grades in case of loss framing.

There is little evidence on framing effects on *school-aged* children. In the educational psychology literature, Kishor and Godfrey (1999) analyze how framing instructions effects academic task completion of third and fourth graders. Pupils were asked to finish an academic task and teachers added information on which consequences—individual or group—students' behavior has. Those consequences were either framed as a gain (*"If you finish these questions ..., there is a 100% chance that your group will receive ..."*) or as a loss *("If you do not finish these questions ..., there is a 100% chance that you will lose...")*. The authors show that task completion rates were significantly higher under all framed instruction conditions.

Closest to my study is the experiment by Levitt et al. (2016) which is the only study—to the best of my knowledge—testing loss framing of an extrinsic reward among *school-aged* children. The authors provide elementary and high school students in Chicago with financial ($10 or $20) and non-financial (a trophy) incentives for a self-improvement in a low-stakes test. These incentives were announced immediately before the test and were presented either as 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.[17] 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 statistical not significant. My study differs in several ways to Levitt et al. (2016): (i) I apply a loss framing on *points in a test* and not on an *extrinsic* reward,[18] (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.

3.3 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

[16] Students had to complete (i) daily quizzes and assignments, (ii) one group project and (iii) three exams including the final exams, each worth 100 points.

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

[18] 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.

May and November 2015, 1.377 pupils in grades 3 and 4 participated.[19] With the semester report in grade 4, 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.[20] Hence, policy interventions to boost pupils' performance in grades 3 and 4 might have long-lasting effects as these grades are important stages for the recommendation decision and promotion within the German school system.

3.3.1 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.[21] 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"*.[22] 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.[23] The problems and the answer options were presented on three question sheets and points could be earned according to the treatment specifications (see Table 3.1). There were five answering possibilities with

[19] Elementary school in Germany runs from grade 1 at the age of 6 to grade 4 at the age of 9 or 10.

[20] 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`).

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

[22] 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.

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

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 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").[24]

3.3.2 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.[25] The grading scheme differed across treatments and was announced to pupils on the testing day shortly before the test started. Hence, this design allows to measure a pure effort effect and no learning because pupils had no time to study after the grading scheme was communicated.[26] Any treatment effects can therefore be attributed to pupils exerting more effort during the test and not to a learning effect—e.g. 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 in case 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 for a correct answer. Likewise pupils in the Control Group, they could earn between 0 and +40 points.

[24] 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 3.5.1.

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

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

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 the Control Group was that pupils started the test with -20 points.[27] Thus, pupils could earn between -20 and +20 points. Usually pupils 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 either only gain points in the Control Group and only lose points in the Loss Treatment. Notice that pupils in 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. Table 3.1 gives an overview of the treatment conditions.

Table 3.1: 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.[28] Blocked on grade level within schools, classes were randomized either 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.[29] The Loss and Negative Treatment were implemented simultaneously for schools participating with three or more classes of the same grade level.

Table 3.6 in Appendix 3.9 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.

[27] Pupils in grades 3 and 4 already learned addition and subtraction with numbers from 0 up to 100. Although they did not learn formally to calculate in the negative range of numbers it is assumable that third and fourth graders understand that having negative points would not result in a good grade.

[28] See Duflo et al. (2007); Bruhn and McKenzie (2009) regarding the rationale for the use of randomization.

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

Table 3.7 in Appendix 3.9 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 effects on pupils' performance and therefore would underestimate positive treatment effects. I therefore take into account differences in teachers' experience in the statistical analysis.

Participants are on average 9.10 years old and have 0.79 older siblings. 48.80% of the pupils 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.[30]

3.3.3 Implementation

The implementation of the experiment is similar to the experiment in Chapter 2. 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 expiration 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 3.9). 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

[30] 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.

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

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.

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.[32] 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). However, the experimental design seems to be superior to these standardized comparative tests as the experiment is conducted in pupils' natural learning environment and pupils get feedback about their test performance the latest after one week. Thus, there are several reasons why pupils should be motivated to put effort into the 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 Chapter 2) and third, giving feedback on performance allows for social comparison within the classroom (Bursztyn and Jensen 2015).[33] Furthermore, there is mixed evidence that performance changes if the test counts towards the course grade. While Baumert and Demmrich (2001) find no differences between high and low-stakes testing with respect to intended and invested effort, Grove and Wasserman (2006) find that grade incentives boosted the exam performance of freshmen but not for older students.[34] Therefore, analyzing grading manipulation in a low-stakes testing environment can shed light on how framing might change performance in a high-stakes testing environment. Nevertheless, it would be interesting to analyze framing effects in high-stakes tests and in long run studies in future research. However, in a first step it was easier to convince teachers to participate in a low-stakes study.

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:

[32] 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.

[33] 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.

[34] Camerer and Hogarth (1999) review the literature on experiments in which the level of financial incentives was varied. They find mixed results of incentives on performance and that the effectiveness of incentives seems to be task dependent.

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.

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."

3.4 Hypothesis

One objective of this paper is to test whether loss framing increases test performance of elementary children. According to prospect theory (Kahneman and Tversky 1979), individuals evaluate a loss approximately twice as much as an equal gain if they are loss averse and therefore choose more often a risky gamble than a sure outcome. In a multiple-choice test, pupils also have the choice between a risky gamble (answering a question) and a sure outcome (omitting a question) if they do not know the answer with certainty. Therefore, if pupils are loss averse, start with the maximum number of points and can only lose points, they should give more answers in the Loss Treatment in order to avoid losing points with certainty. The underlying assumption is that pupils' reference point is their current asset (+40 points) and due to loss aversion change their behavior compared to the Control Group. However, if pupils are not loss averse or their reference point does not change to the new endowment, there should be no difference between the Control Group and the Loss Treatment. Nevertheless, informed by previous research, I hypothesize that pupils are loss averse, adjust their reference point to the new endowment and therefore choose more often the risky option, i.e. increase the quantity of answers.

Hypothesis 1 *The number of answered questions in the Loss Treatment is higher than in the Control Group.*

The Negative Treatment and the Control Group differ only with respect to their initial endowment of points. This means, the point scale is shifted downwards which could—according to prospect theory—effect pupils' performance in two ways: First, they could adjust to the incurred loss of -20 points and accept this endowment as their new reference point. In this case, earning points is in the domain of gains and performance should not differ from the Control Group. Second, pupils do not immediately adjusted to the new endowment and their reference point is at 0 points—the "traditional" starting point. In this case, pupils would face a negative discrepancy between the reference point and their current endowment. Hence, they are likely to code their situation as a loss which could result in an increase in their performance. If this would be indeed the case, pupils' behavior should be changed by the same mechanism (loss aversion) as in the Loss Treatment. This means, pupils would also chose more often the gamble. However, pupils in the Negative Treatment might also increase their performance if they adjust their reference point to the new endowment. The Negative Treatment increases the salience of the "passing" threshold and therefore sets an intermediate goal at 0 points, whereas in the Control Group pupils' goal is at +40 points. Hence, pupils in the Negative Treatment are closer to their (intermediate) goal and due to diminishing sensitivity of the value function increase their test performance. This increase can be reached by answering more questions, answering questions more accurately or a mixture of both. Moreover, pupils could also adjust to the incurred loss and simply have more pessimistic beliefs about the grade they get if they score negatively. Thus, I expect that pupils in the Negative Treatment perform better in the test than pupils in the Control Group.[35]

Hypothesis 2 *Pupils in the Negative Treatment perform better in the test compared to pupils in the Control Group.*

It is of crucial importance to inform policy makers and educators about heterogeneous framing effects to know for whom loss framing potentially works (negatively). There is evidence that pupils who differ in their cognitive ability also differ in risk preferences, i.e. that cognitive ability and risk aversion are negatively related (Benjamin et al. 2013; Dohmen et al. 2010; Burks et al. 2009) and Frederick (2005) shows

[35] Whether the Negative Treatment has long run effects on pupils performance cannot be answered in this study. It might be that the negative endowment of points results only in short run effects if pupils learn to adjust their reference points to the incurred loss in repeated interventions. However, short run interventions can give valuable insights on how long run studies might work. If the Negative Treatment does not motivate pupils in the short run then it is also unlikely that motivation would increase in repeated interactions.

that individuals who score high on a cognitive reflection test (CRT) are more risk-seeking in gain domains and less risk-seeking in loss domains than individuals scoring low in the CRT.[36] Low-ability pupils could therefore be more sensitive to losses than high-ability pupils. Hence, if loss aversion is assumed to be the mechanism boosting performance, the difference in performance between low-ability pupils in the Loss Treatment and low-ability pupils in the Control Group should be larger than the difference between high-ability pupils in the Loss Treatment and high-ability pupils in the Control Group (see also Imas et al. 2016 on sensitivity to loss averion).

Hypothesis 3 *Low-ability pupils are more sensitive to losses which leads to larger differences in performance compared to high-ability pupils.*

3.5 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 3.6 are gender, parents' education and a dummy whether pupils are in grade 3 or 4. 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.[37] 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 4

[36] Andersson et al. (2016) report evidence that the negative relation of cognitive ability and risk aversion may be spurious as they find suggestive evidence that cognitive ability is related to random decision making rather than to risk preferences.

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

before the summer break are one year ahead in the teaching material than pupils in grade 4 after the summer break.

Table 3.2 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.[38] 1.333 observations were included in the final analysis; 44 observations were dropped because of missing values.[39]

Table 3.2: Comparison of Characteristics: Experiment vs. North Rhine-Westphalia (in percent)

	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.

3.5.1 Attrition

Parents had to give their consent that their child is allowed to participate in the experiment and that teachers are allowed to pass on pupils' test as well as midterm grades to the researcher.[40] 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 at the testing day, (ii) pupils have lost or forgotten the signed consent form, (iii) parents forgot to timely sign the consent form but actually agreed or (iv) parents intentionally did

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

[39] 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.

[40] 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`).

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 3.8 in Appendix 3.9 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, this small difference 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.

3.6 Experimental Results

The result section is organized in the following way. First, the effectiveness of framing on the number of correct answers is analyzed using Poisson regression models (ordinary least square regressions are presented in Table 3.15 in Appendix 3.9). Thereafter, treatment effect estimates are presented for the number of omitted questions and total points using 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.

I first analyze treatment effect estimates for the number of correct answer instead of the number of total points because teachers are likely to be more interested in the former. 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.

3.6.1 Framing and Test Performance

The outcome variable of interest (for the moment) is the number of correct answers in the test and represents count data. 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:

$$E(NumCorrect_i) = m\left(\beta_0 + \beta_1 Treatment_i + \beta_2 Midterm_i + \gamma P_i + \mu C_i + \delta School_i\right) \tag{3.1}$$

m(.) is the mean function of the Poisson model. $NumCorrect_i$ is the number of correctly answered questions 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 3.5) 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 3.15 in Appendix 3.9).

Table 3.3 presents estimates of the average treatment effects for the Loss Treatment and Negative Treatment. The dependent variable is the number of correct answers 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.[42]

Pupils in the Loss Treatment as well as pupils in the Negative Treatment increase, as expected, 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.

Result 1 *Loss framing and a negative endowment increase significantly the number of correctly solved questions.*

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

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

Table 3.3: 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 3.15 in Appendix 3.9).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

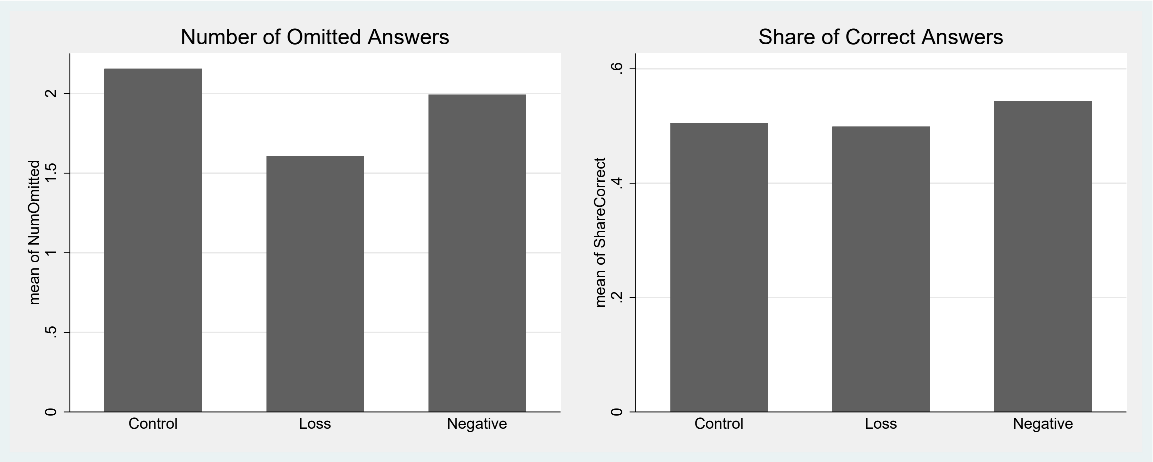
[43] Risky multiple-choice question refers to a test question where the answer is unknown and thus answering this question is a decision under uncertainty.

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 between the Control Group and the treatments groups would be an indication of a change in risk-seeking behavior. 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 3.1—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 3.1 shows the average number of omitted questions (left) and the average share of correct answers (right) of pupils by treatments.

Figure 3.1: 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 3.1. 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 3.4 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).

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

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 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. Nevertheless, there is *suggestive* evidence that treatments increase overall performance as coefficients on the number of total points are positive (as expected); however, this result is not definitive.

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 2 *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 3.4: Treatment Effects - All Outcome Variables

	(1) *Correct Answers*	(2) *Omitted Answers*	(3) *Share Correct Answers*	(4) *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 3.15 in Appendix 3.9) and estimation of treatment effects without any controls except including school fixed effects (see Table 3.12 in Appendix 3.9) show similar results.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.6.2 Who can be Framed?

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.

[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.

Table 3.5 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 3.16 in Appendix 3.9 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]

[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 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`).

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 3 *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 3.5: 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

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 3.13 in Appendix 3.9 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 3.2 in Appendix 3.9). 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) and discussed in more detail in *Chapter 4*. 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 4 *There are no detectable heterogeneous gender effects on performance when the grading scheme is manipulated.*

3.7 Discussion

Here, I want to address three further questions: First, do pupils in the Loss Treatment answer marginally more difficult questions? Second, do pupils change their answering behavior when they reach the threshold of "passing"? Third, which questions are considered as difficult and do pupils in the Loss Treatment answer strategically by choosing more easy 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 3.12 in Appendix 3.9 shows kernel density estimates for the number of points in the test for the Control Group and Negative Treatment.[49] 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 3.3 in Appendix 3.9. In Figure 3.3 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 3.4 in Appendix 3.9 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?

[49] Further kernel density estimates on the number of points and number of correct answers can be found in Appendix 3.9

Table 3.14 in Appendix 3.9 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, 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. 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.

3.8 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 pupils in both treatment groups answer significantly more questions correctly compared to pupils that are graded "traditionally". These improvements are 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—same amount of answered questions but an increase in the share of correct answers.[50] Moreover, pupils can be differentiated by their ability—as measured by their past midterm grades. Both treatments work equally good to increase performance of high-ability pupils. In contrast, performance is significantly decreased for low-performers in the Loss Treatment but not for low-performers in the Negative Treatment.

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 the field. 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 of. Moreover, the experimental design allows to isolate the effort effect from a learning effect, showing

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

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.

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). 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, e.g. 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.

3.9 Appendix

Randomization Table

Table 3.6: Sample Size by Gender, Grade 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 3.7: 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

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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Attrition

Table 3.8: Attrition 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.

Estimation Tables

Table 3.9: 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 3.10: 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 3.11: Treatment Effects - Total Points in Test

	(1)	(2)	(3)	(4)
Treatments				
Loss	-0.073	-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.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.12: Treatment Effects without Control Variables- Correct, Omitted, Share and Points

	(1) *Correct Answers*	(2) *Omitted Answers*	(3) *Share Correct Answers*	(4) *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 3.13: 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 3.14: 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. Standard errors are reported in parentheses and clustered on classroom-level. The number of clusters is 71.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Robustness Checks

Table 3.15: 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.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

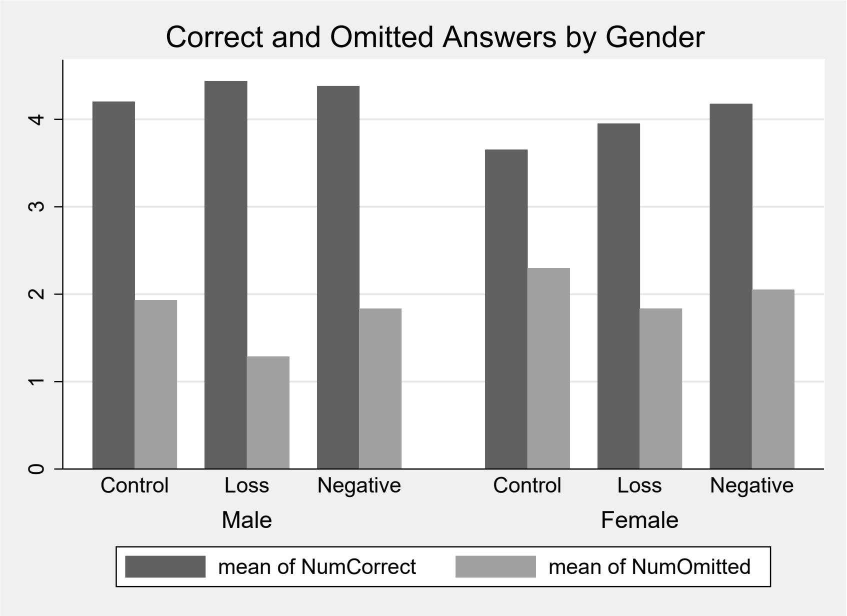
Table 3.16: Treatment Effects by Midterm Grade

	(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 3.5 in Section 3.6.2, 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 3.6.2, 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

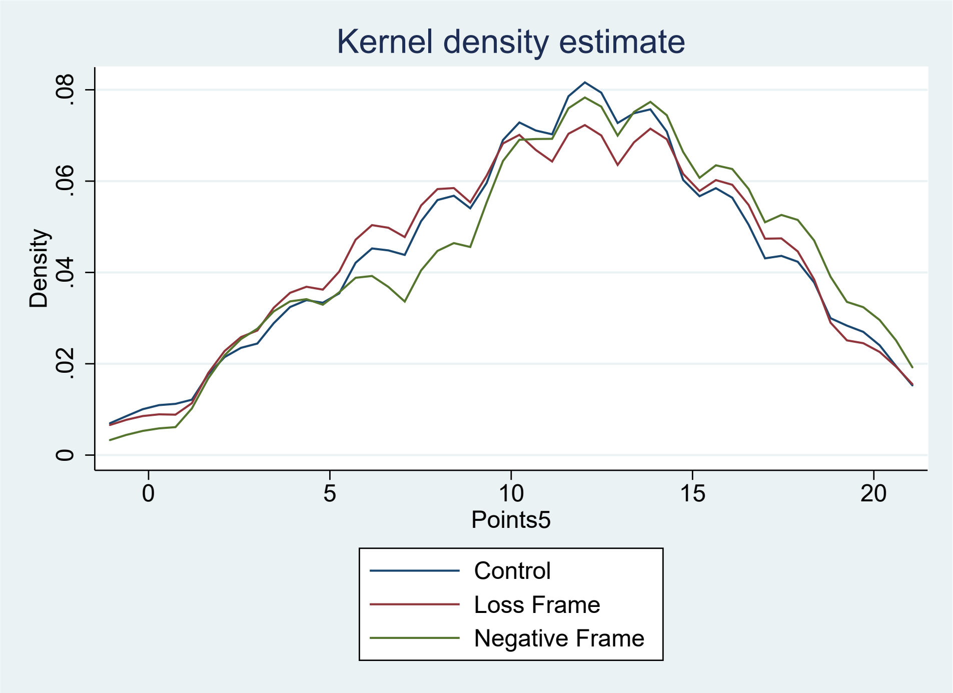
Figures

Figure 3.2: Average Number of Omitted Answers and Share of Correct Answers by Gender



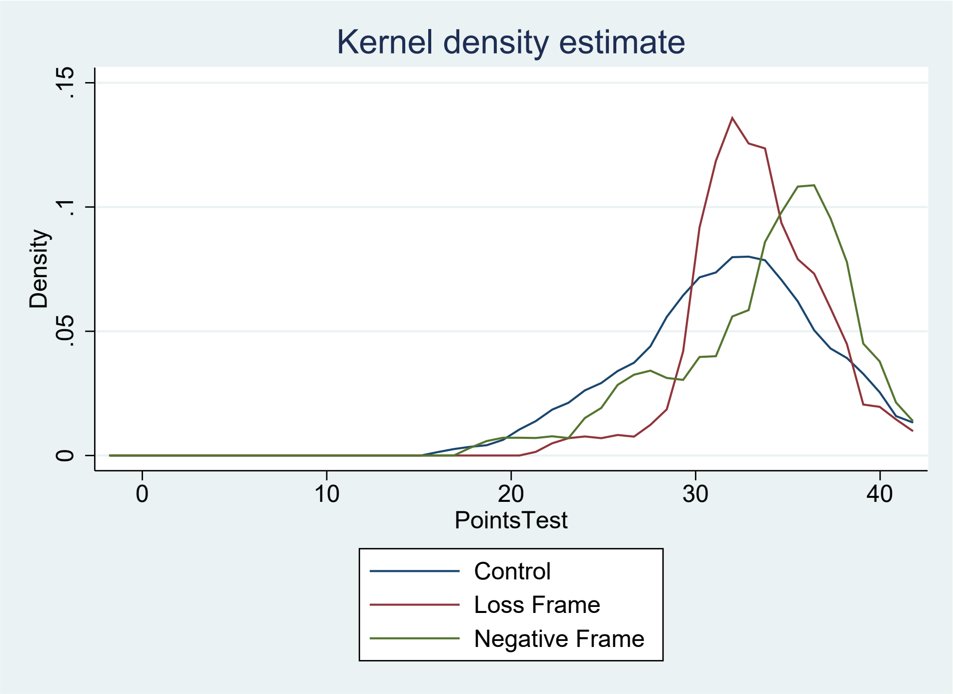
Note: This figure reports the average number of correct and omitted answers separately for boys and girls.

Figure 3.3: 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 3.4: 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.

Instructions, Questionnaire and Consent Form

Instruction for Teacher

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 3.3.

Figure 3.5: 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 3.6: 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.

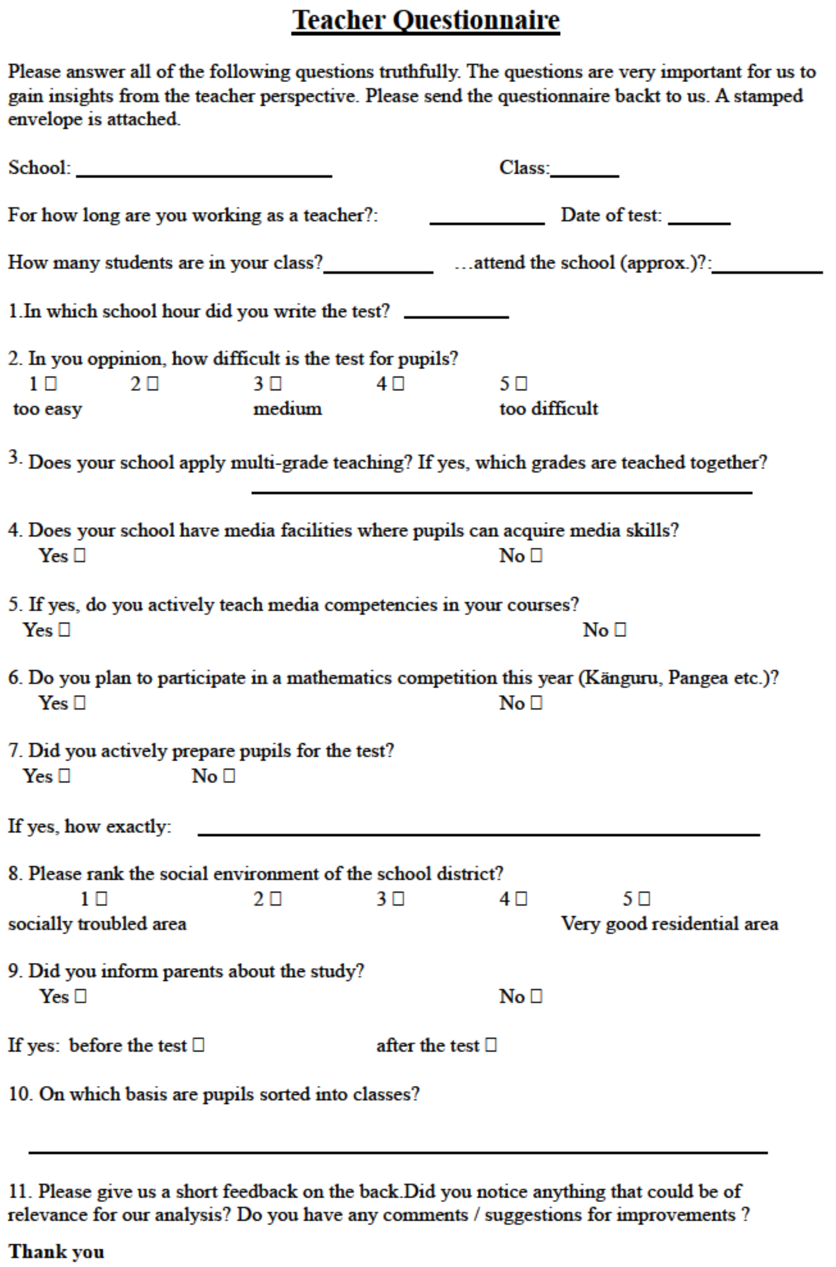
5. While the test is ongoing, please write down the corresponding name for each Test-ID number (upper left corner on the first page of the test) on a sheet of paper. For this, you could also use a class list. This sheet serves as an "encryption key" which you do not send back to us and keep for yourself. 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.

Teacher and Student Questionnaire

Figure 3.7: 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 □

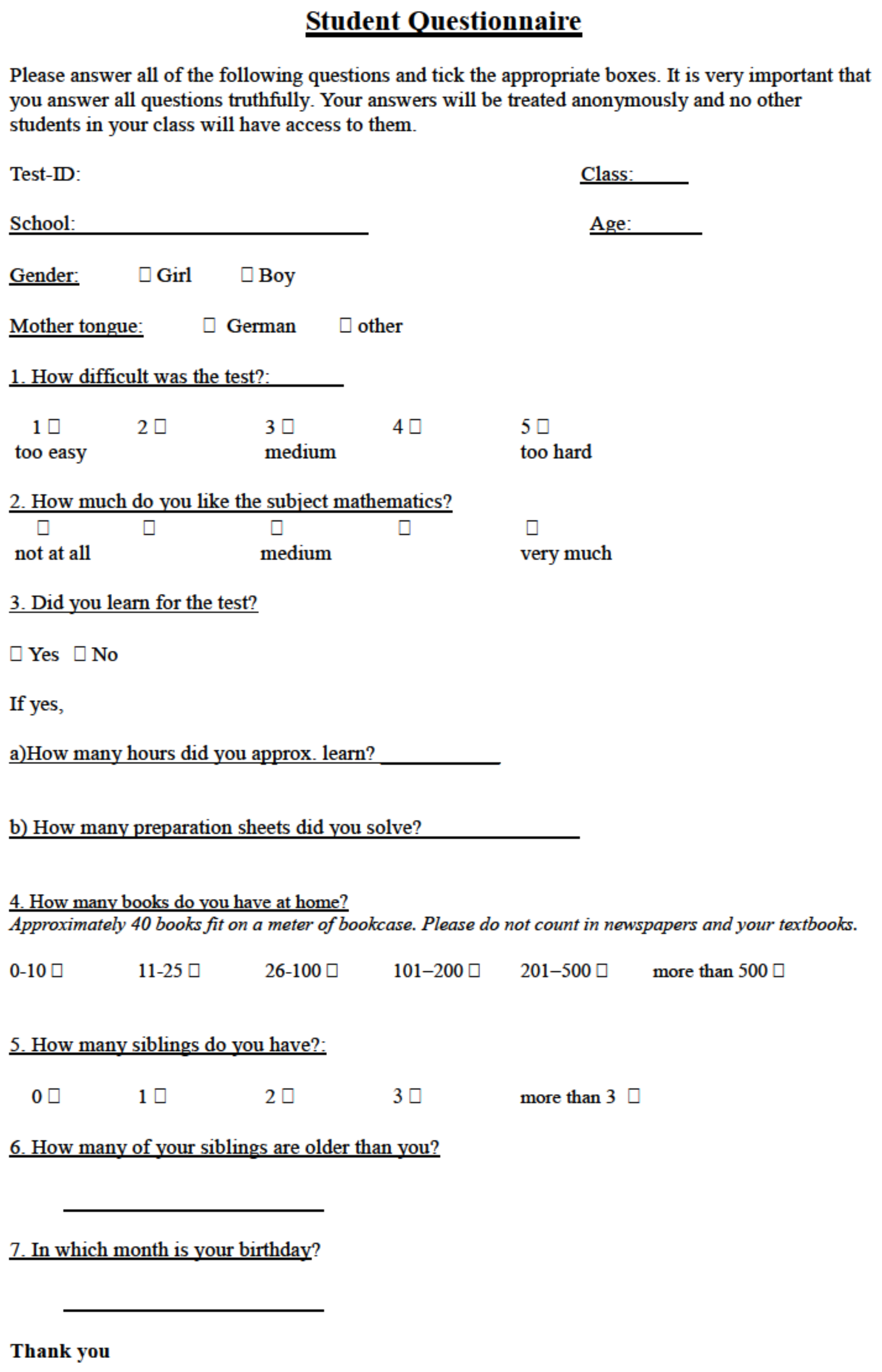
10. On which basis are pupils sorted into classes?

____

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 3.8: 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

Consent Form

Figure 3.9: 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,

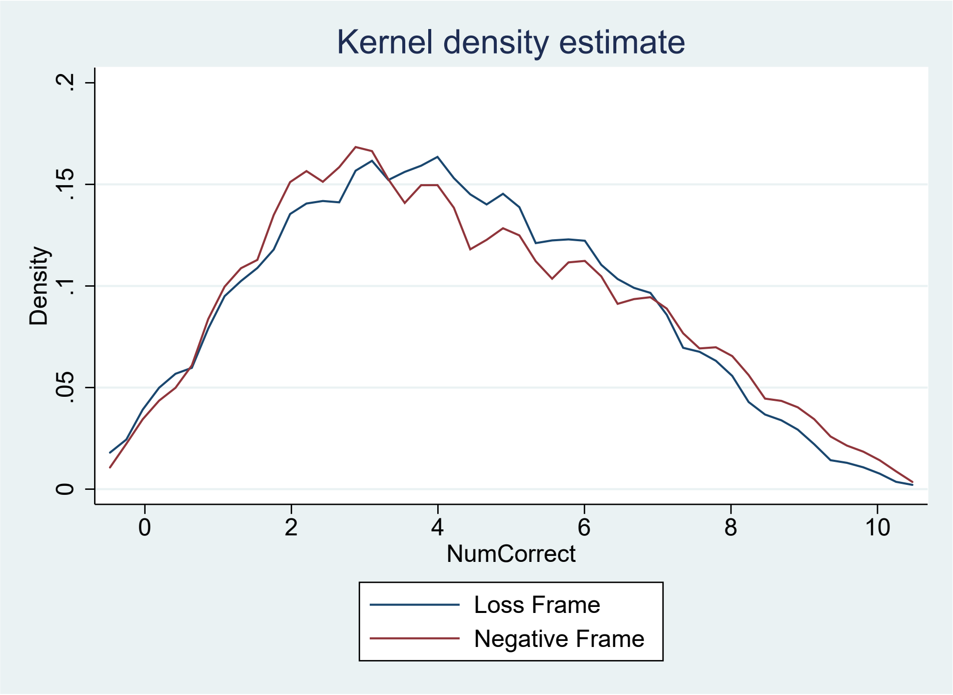
Declaration of Consent for study participation

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. 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)

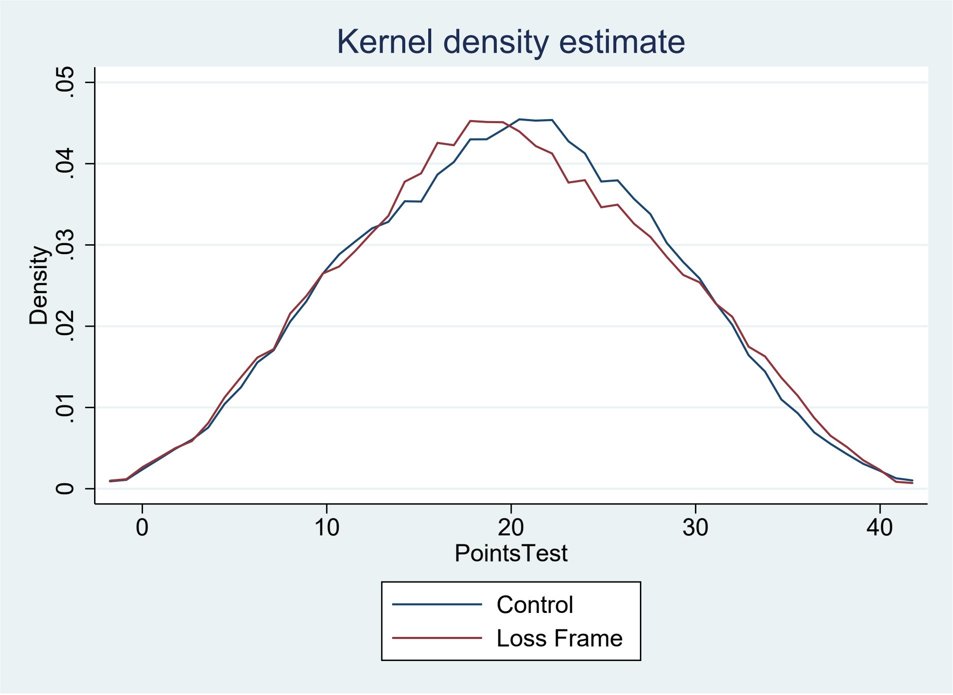
Kernel density plots by Treatment

Figure 3.10: Correct Answers: Loss Treatment vs. Negative Treatment



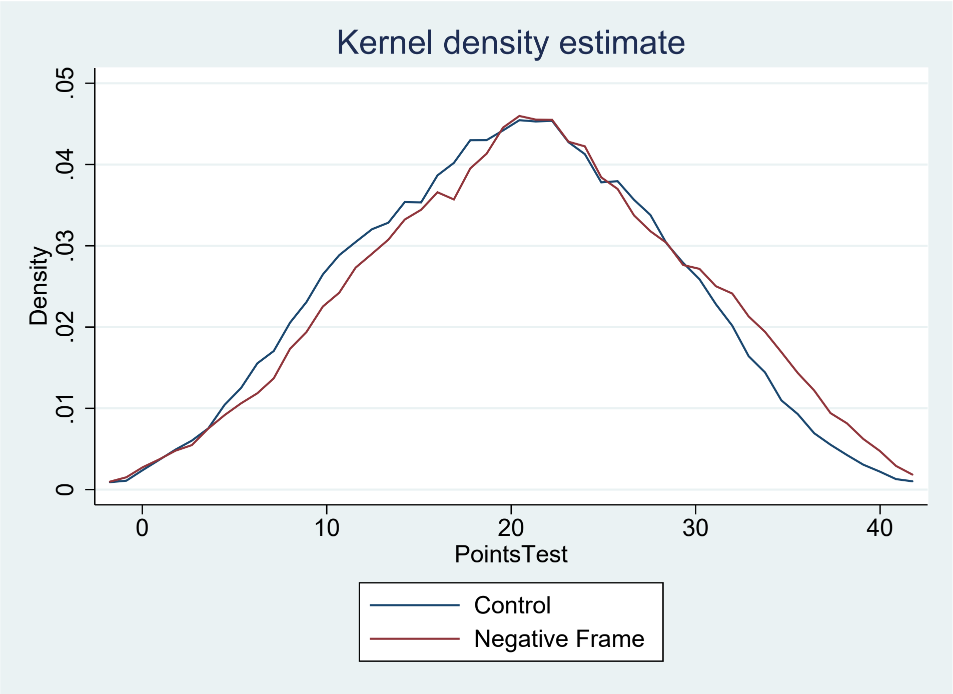
Note: This Figure presents kernel density estimates for the number of correct answers for the Loss Treatment and the Negative Treatment.

Figure 3.11: Points: Control vs. Loss Treatment



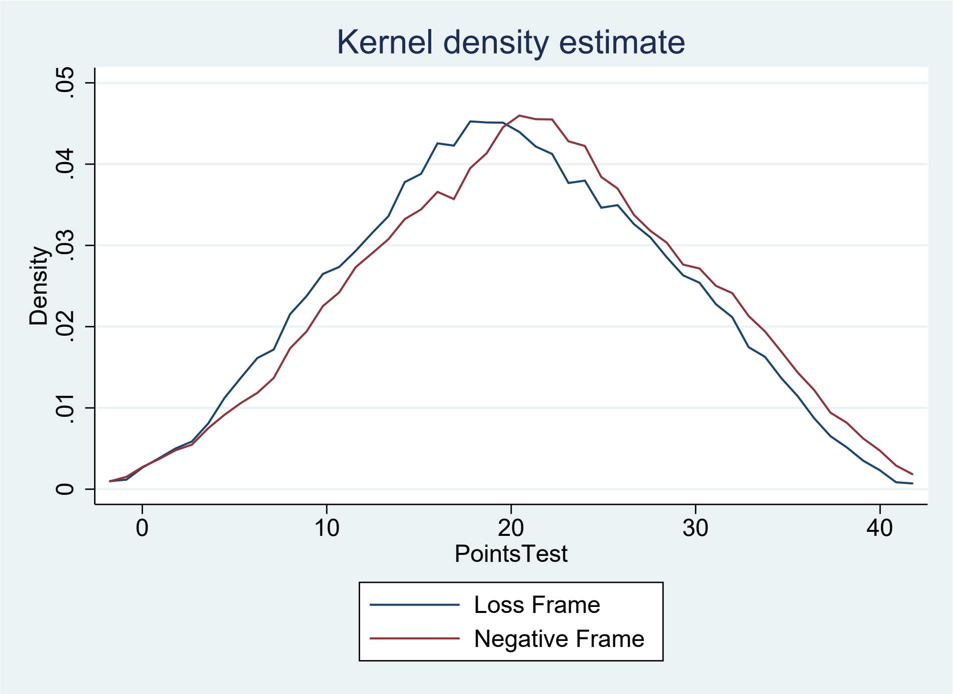
Note: This Figure presents kernel density estimates for the number of points reached in the test for the Control Group and the Loss Treatment.

Figure 3.12: Points: Control vs. Negative Treatment



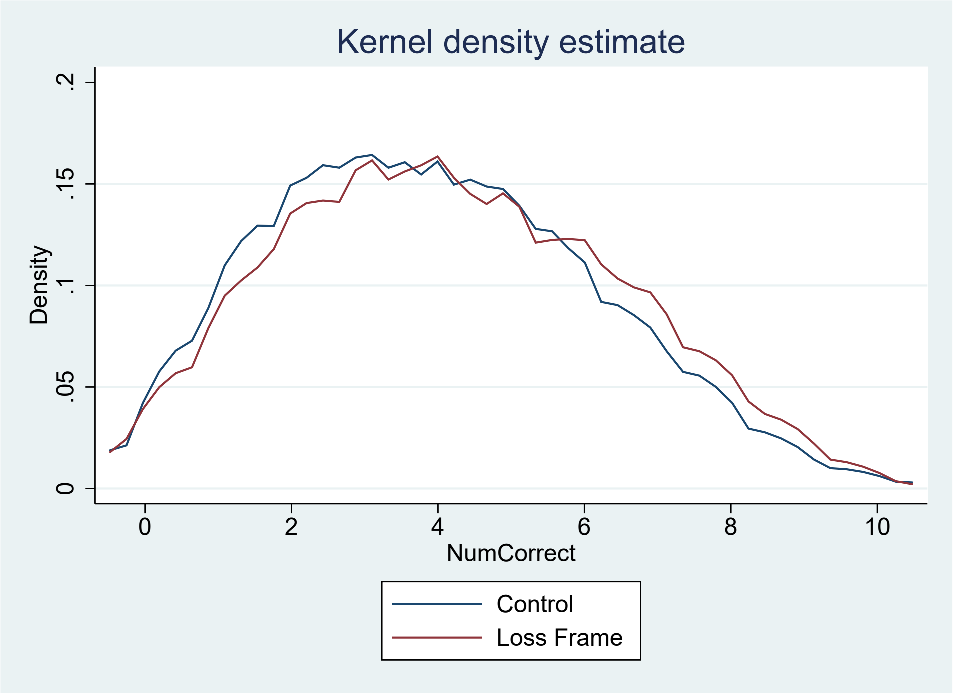
Note: This Figure presents kernel density estimates for the number of points reached in the test for the Control Group and the Negative Treatment.

Figure 3.13: Points: Loss Treatment vs. Negative Treatment



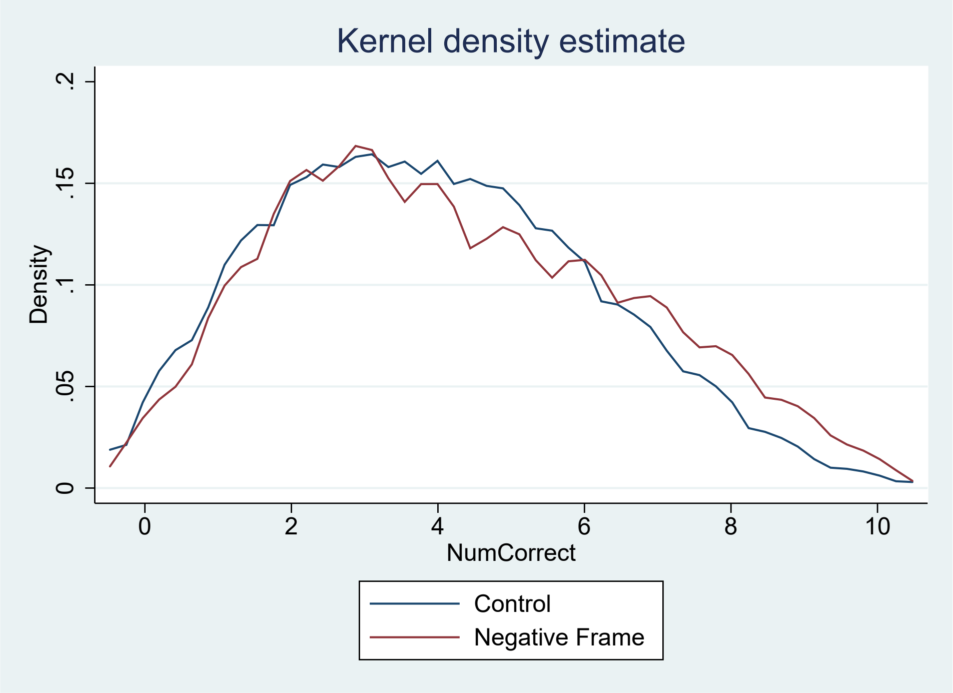
Note: This Figure presents kernel density estimates for the number of points reached in the test for the Loss Treatment and the Negative Treatment.

Figure 3.14: 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 3.15: 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.

Chapter 4

Answering Strategies in Multiple-Choice Tests - Differences by School Types and Gender?

Co-authored with Gerhard Riener

Contributions of Valentin Wagner

- Development of research idea and literature review
- Establishing contact to schools and preparatory talks
- Design of experiment and expiration
- Data preparation, descriptive analysis and graphs
- Treatment effect estimation and robustness checks
- Textual contributions in all sections of the paper

____

Gerhard Riener

4.1 Introduction

In recent years there has been an increase in the inequality of wages across groups within many societies which is largely driven by the returns to formal education (Lemieux 2006). Moreover, there is an ever increasing wage gap between socially disadvantaged groups (Autor et al. 2008) and although the gender gap in educational achievement has been reduced or reversed in most subjects (see Niederle and Vesterlund 2010; Goldin et al. 2006; Duckworth and Seligman 2006; Hyde and Mertz 2009; Fortin et al. 2015)[1], women still earn on average around 16% less per hour than men in the EU (The European Commission 2014) and around 18% less in the US (The US Bureau of Labour Statistics 2014).[2] The reasons for these wage and education gaps are complex and manifold.[3] The access to higher education is one important prerequisite for later employment possibilities and wages and is determined *inter alia* by university entrance exams in many countries.[4] These exams often use multiple-choice testing formats—especially in the US—because it is considered as efficient, it allows for large scale testing and for a broad coverage of content (Frederiksen 1984).[5] Nevertheless, multiple-choice testing formats are not without problems if they favor answering strategies of certain groups in the population.

The analysis of gender differences in standardized multiple-choice tests with respect to *performance* (Ors et al. 2013; Jurajda and Münich 2011) or *skipping test items* (Pekkarinen 2015; Akyol et al. 2016; Ben-Shakhar and Sinai 1991) has therefore received increasing attention. Recent experiments have identified guessing—women ten to skip more test items than men—as one reason for men outperforming women (Pekkarinen 2015; Baldiga 2014), but as promotion within the educational system should depend on actual knowledge and not on how knowledge is assessed this poses a challenge for general multiple-choice tests. The negative effect of skipping test items on performance has also been recognized by the College Board which recently (March 2016) redesigned the scoring rules of the SAT. The old scoring rule—students get $\frac{1}{4}$ point deducted for incorrect answers—has been changed to a

[1] Goldin et al. (2006) show that females gained about 0.17 of a standard deviation from 1972 to 1992 in standardized math tests in the US and Hyde et al. (2008) show that gender differences in mathematics skills are close to zero for grades 2–11. Hyde et al. (2008) analyze scores on the National Assessment of Educational Progress (NAEP) of about 7 million eight graders of ten states in the US. In contrast using data from the Early Childhood Longitudinal Study Kindergarten Cohort, Fryer and Levitt (2010) find that there are no mean gender differences upon entry to school in math standardized test scores, but that girls lose more than two-tenths of a standard deviation relative to boys over the first six years of school.

[2] Furthermore, the wage gap between skilled and unskilled population groups has been rising since—at least—the 1970s (see Marquis et al. 2014 and the literature mentioned therein).

[3] According to the European Commission possible explanation could be inter alia discrimination in the workplace, different jobs in different sectors (STEM fields), undervaluing of women's work and skills along with women's under-representation in senior and leadership positions.

[4] In the US, the weekly earnings of women having only a high school diploma represented 83% of the earnings of women with an associate's degree and 55% of the earnings of women with a bachelor's degree or higher (The US Bureau of Labour Statistics 2014).

[5] Also in Germany multiple-choice questions constitute an important testing format in centralized comparison tests (VERA, PISA, TIMSS) and in university exams. Furthermore, testing of cognitive knowledge also predicts and correlates well with overall competence and performance (McCoubrie 2004).

"rights-only" scoring method. Under the new scoring rule students receive one point for each correct answer and each incorrect answer receives zero points "*to encourage students to give the best answer they have for every question without fear of being penalized for making their best effort*" (The College Board, 2016).[6] Structural biases in multiple-choice testing would therefore challenge the use of this testing format and understanding the underlying causes is important, in particular if differences are driven due to a higher willingness to skip questions and not due to differences in ability.[7]

Although any testing format is advantageous for some and less advantageous for others (e.g. oral exams could favor extroverts; open-ended questions could favor females. . .) taking all characteristics of the general population into account in the design of tests for promotion within an educational system and its consequences for intergenerational persistence of educational differences is important to (at least) not increase educational inequalities and to develop an institutional setting for equal opportunities in education. So far, researchers have mainly focused on gender differences in university entrance examinations and hence relied on student subjects which does not allow to analyze differences in skipping for a more heterogeneous population. However, standardized test scores are used for placements and admissions at nearly every level of schooling (Baldiga 2014) and it is therefore important to know at which stage, respectively age, gender differences emerge and how these differences could be mitigated. What has not been investigated so far, is whether social background matters in test answering strategies and at which stage of the education production function gender gaps occur.

In this paper, we analyze how pupils answer multiple-choice questions, whether differences in answering strategies exist and whether this differences depend on pupils' socio-economic background, gender and school grade. To do so, we designed a field experiment among fifth and sixth graders—right after the first tracking decision—in 25 secondary schools of both the academic and the vocational track in Germany. Our test questions originate from a German-wide mathematics competition test (*Känguru-Wettbewerb*[8]) where we have reliable data on item difficulty.[9] We vary test item difficulty within subjects and increase the attractiveness of difficult questions. The expected value from randomly guessing is negative for easy questions, zero for medium questions but positive to answer difficult questions. Skipping difficult questions is therefore never an optimal strategy. This gives us a strong test of the persistence of the gender gap for difficult items when answering is made more attractive. Moreover, we use external incentives to increase the stakes of the test; pupils received non-monetary rewards for improving over their own previous mathematics results. This allows us to test whether the gender gap widens or can be closed if item difficulty is less salient by setting the focus on winning a prize. Conditional on their ability, females are found to be less willing to contribute ideas

[6] See https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf.

[7] Other reasons for the answering gap in (mathematical) multiple-choice tests could be women's retention for competitive settings (Niederle and Vesterlund 2010; 2007) or the stereotyping that women perform worse in mathematics than man.

[8] http://www.mathe-kaenguru.de/wettbewerb/.

[9] We could not vary the order of item difficulty—easy to difficult—as this was also predetermined by the *Känguru-Wettbewerb* and a prerequisite of schools to participate in the study.

in areas that are stereotypically outside of the gender's domain (Coffman 2014). Hence, by setting the focus on winning a prize, based on self-improvement, could reduce or close gender differences in guessing. Finally, we complement our field experiment with the data from the official *Känguru-Wettbewerb* from 2013 to compare our experimental results with a broad sample of over 780.000 German pupils from grades 3 to 12 and to test for gender differences in guessing over school grades.

We implement our study within the German school system because it is particularly well-suited to study the role of socio-economic background on answering strategies: the transition decision in Germany depends heavily on parents' socio-economic background (Dustmann 2004), there is little mobility across school types and pupils of the same age are segmented either into schools which prepare for vocational jobs ("*Vocational School*") or higher education *("High School*"). Hence pupils' social background and their intellectual ability is measured by the school type. Furthermore, the tracking decision takes place early at the age of ten. At this early stage, we can identify whether differences are due to selection rather than caused by the school types (i.e. because higher achieving schools may better prepare for test situations and multiple-choice testing).

Overall, our results suggest large differences between school types. Pupils in High School tend to skip more questions than pupils in Vocational School. Nevertheless, pupils in High School score higher in the test than pupils in Vocational School, as the share of correctly answered questions is higher, suggesting that pupils differ in their answering strategy. To the best of our knowledge, this is the first study that compares the answering behavior between pupils who differ with respect to their socio-economic background. Moreover, there are only a few studies using data from framed field experiments (Espinosa and Gardeazabal (2013) on second year undergraduate students and Baldiga (2014) on participants at Harvard Business School) and no randomized field experiment has been conducted so far in secondary education. Moreover, we are the first ones to examine gender differences for (almost) all grade and age levels within a school system. On grade levels, we find that boys as well as girls in higher grades tend to skip more questions than boys and girls in lower grades and that gender differences tend to increase from grade 3 to grade 12. Moreover, in our—non-incentivized—baseline treatment, we find that girls in High School skip more questions than boys if test items are difficult. This gender differences are not found in Vocational School. Hence, our findings in High School are in line with the findings in the literature but furthermore show the importance to consider all social levels of a population. Interestingly, gender differences in skipping are not detectable anymore and small in size if pupils can win a reward.

Recent literature documents that girls outperform boys in terms of GPA but that boys still perform better on standardized tests (Saygin 2014; Fortin et al. 2015; Goldin et al. 2006; Duckworth and Seligman 2006).[10] One explanation could be

[10] The literature on multiple-choice testing has analyzed a variety of issues such as validity, reliability, discriminative power, gender bias, cultural bias, objectivity, guessing and cheating. Gafni and Melamed (1994) for example analyze the influence of different cultural backgrounds on the guessing behavior. They find that people with differing cultural backgrounds, as well as male and females, differ in their tendency to guess. Tamir (1993) looks at the difference of a positive and negative item mode. In the positive item mode individuals have to identify the unique correct answer whereas in the negative item mode they have to identify the unique wrong answer. The

that boys outperform girls when facing novel problems presented in standardized tests and that girls are more confident answering questions about familiar material (Kimball 1989; Loewen et al. 1988). Another explanation could be the way pupils are allowed to answer test questions—open-ended question vs multiple-choice testing—or gender differences could exists due to increasing pressure (Azmat et al. 2016). Azmat et al. (2016) show that male and female high school students react differently to tests with varying stakes. Female students outperform male students in low-stakes tests but to a smaller extent in high-stakes tests. There seems to be a general agreement in the education literature that multiple-choice questions tend to favor males over females (Ben-Shakhar and Sinai 1991; Bolger and Kellaghan 1990; Stumpf and Stanley 1996) and similar findings have been confirmed in recent economic studies (Baldiga 2014; Jurajda and Münich 2011; Tannenbaum 2012; Espinosa and Gardeazabal 2013). Another cause that has been identified in recent economic studies is a gender gap in the willingness to guess in multiple-choice tests (Pekkarinen 2015; Baldiga 2014; Akyol et al. 2016; Tannenbaum 2012). Pekkarinen (2015) analyzes the performance and number of skipped questions in the joint entrance examination of Finnish universities. The author shows that women perform worse than men in the entrance exam and are less likely to gain entry. This is because women skipped more questions than men and therefore deviated more than men from the number of items that would maximize the predicted probability of entry. Espinosa and Gardeazabal (2013) conducted a field experiment and rewarded skipped questions with a positive number of points (0.5 points). In total they tested three scoring rules (i) penalty for incorrect answers (ii) normalized penalty for incorrect answers and (iii) normalized reward for omission. As expected Espinosa and Gardeazabal (2013) find that being rewarded for omissions tends to increase the number of omissions and that the accumulated score in previous exams, item difficulty, other unobserved characteristics of the exam and gender are significant determinants of omissions—males tend to be non-omitters.

Closest to our study is the experiment by Baldiga (2014). The author analyzes whether women are more likely than men to skip questions rather than to guess for questions of the SAT II subject test. The size of the penalty for wrong answers was varied across subjects and treatments.[11] Baldiga (2014) designed an unframed and SAT-framed version of the treatments. The SAT frame was designed to create a high-pressure environment. Additionally, subjects' confidence in knowledge of the material, differences in risk preferences, and differential responses to high pressure testing environments was measured and related to the guessing behavior. Baldiga (2014) finds no gender gap in the willingness to guess in the no punishment treatment—subjects answers all questions. If there is a penalty for a wrong answer, women answer fewer questions than men. Furthermore, gender remains a significant

author finds that there is no difference between the positive item and negative item modes for low cognitive questions. However, individuals perform better in the positive mode compared to the negative mode in questions which require high level reasoning.

[11] In the low penalty treatment subjects received one point for a correct answer, no point for an omitted question and one quarter of points was deducted for a wrong answer. In the no punishment treatment wrong answers were not penalized for incorrect answers. Each question contained four possible answers so that randomly guessing of a risk-neutral subject had a positive value in both treatments.

predictor of skipped questions, even after controlling for knowledge of the material, levels of confidence, and risk preferences.[12] Importantly, Baldiga (2014) finds that test takers who skip questions do significantly worse on the test.

Our contribution to the literature is therefore threefold. First, we examine answering strategies for pupils who differ in their social background and intellectual ability as measured by the school type. Second, we investigate whether gender differences in skipping items exist in all school grades of the school system. Third, we vary the degree of item difficulty within subjects. Thus, we can analyze if the skipping behavior is dependent on task difficulty.

The remainder of this paper is organized as follows. In Section 4.2 we give background information on the design of the multiple-choice test, the experimental setup and the *Känguru-Wettbewerb 2013*. In Section 4.3 we present our data and descriptive statistics. Section 4.4 summarizes our results which are discussed in Section 4.5. Section 4.6 concludes. In Sections 4.2 - 4.4 we distinguish between experimental and field data.

4.2 Data Sources and Experimental Procedure

The institutional setting, the experimental design and the multiple-choice test of Chapter 2 and Chapter 4 are identical and therefore not described in detail in this Chapter. The experimental design was chosen to shed light on two questions: (i) whether differences in answering strategies in multiple-choice tests exist (focus of this Chapter) and (ii) to which target audience do pupils want to reveal their educational performance (focus of Chapter 2). In the following the details of the multiple-choice test which were not yet described in Chapter 2 (mainly the expected number of points from guessing in each section) are presented briefly and then the structure of the field data is explained.

Design of the Multiple-Choice Test

Känguru-Wettbewerb We received permission to use questions from a mathematics competition test (*Känguru-Wettbewerb*) that is administered throughout Germany and in over 50 other countries.[13] The aim of the *Känguru-Wettbewerb* test is the popularization of the subject mathematics and was carried out in 2014 for the 20th time in Germany. By solving the test questions, the joy of mathematical

[12] Baldiga (2014) also designed a third (high penalty) treatment where one point was deducted for a wrong answer which means that randomly guessing had a negative value. In contrast to the previous two treatments, the high penalty treatment was not framed as an SAT. In the high penalty treatment, neither men nor women skip significantly more questions. This is in contradiction to the findings of Burns et al. (2012) who find that girls are substantially more likely than boys to skip questions on a multiple-choice test as the size of the penalty is increased. Burns et al. (2012) deducted 0, -0.5 or -1 points for an incorrect answer. However, in all cases the expected value of guessing remained positive However, Baldiga (2014) stresses that the small sample size, particularly among men, requires to use caution in interpreting the results of the high penalty treatment.

[13] For further information see `http://www.mathe-kaenguru.de/` or `http://www.mathkangaroo.org/mk/default.html` for the USA.

thinking and working shall be awakened and supported. Participation in the competition test is voluntary for all students of all school types in grades 3–13. Each grade level receives age-appropriate tasks which have to be solved within 75 minutes. All tests in each school grade consist of three difficulty levels that are rated with 3, 4 or 5 points for each correct answer. Tasks are designed such that for some questions basic knowledge and for other questions deeper understanding is sufficient for the solution. All tasks have in common to train mathematical working methods in an enjoyable way.

Multiple-choice test of the experiment The design of the test was predetermined by the institutional setting of the experiment. We could not freely vary the ordering of the difficulty level of the questions—hard questions first or hard questions last—because one prerequisite by schools to participate in the experiment was to not change the setting as well as to use old questions of the *Känguru-Wettbewerb* test.

The problems and the possible choices were presented on three question sheets and pupils received 3 (*easy section*), 4 (*medium section*) or 5 (*difficult section*) points for correct answers. In the *easy* and *medium* sections five questions had to be answered, the *difficult* section consisted of four questions. In Vocational Schools, the *medium* and *difficult* section consisted of two questions which belong originally to the respective difficulty category. The remaining questions in these sections were taken from the easy section in the original *Känguru-Wettbewerb* test. This was necessary to fulfill teacher prerequisites and to account for pupils' lower ability compared to pupils in High School.[14] In our analysis, we therefore exclude the easy questions in the *medium* and *difficult* section in Vocational Schools. There were five answering possibilities with only one correct answer per question, and pupils had to mark their answers on the same sheet. The amount of points for a correct answer was clearly presented to pupils at the beginning of each section. We deducted one point for each wrong answer regardless of the difficulty level. Skipping an answer counted zero points. Thus, the amount of expected points from guessing for a risk-neutral subject varies with each sections:

$$E(points) = \begin{cases} 3 \times \frac{1}{5} - 1 \times \frac{4}{5} = -\frac{1}{5} & \text{“\textit{easy section}”} \\ 4 \times \frac{1}{5} - 1 \times \frac{4}{5} = 0 & \text{“\textit{medium section}”} \\ 5 \times \frac{1}{5} - 1 \times \frac{4}{5} = \frac{1}{5} & \text{“\textit{difficult section}”} \end{cases} \tag{4.1}$$

Therefore, a risk-neutral subject who does not know the answer should always skip items in the easy section and is indifferent between guessing and skipping in the medium section. Skipping answers is never the best strategy in the difficult section as randomly guessing gives a positive expected value.[15]

[14] Our aim was to design a test in which pupils had enough time to answer all questions. If the test would have been designed too hard for pupils in Vocational School, it would not be possible to identify whether questions were omitted because pupils did not want to answer them or because they did not have enough time.

[15] Note that also for a risk-averse subject guessing is more attractive in the difficult section than in the medium and easy section.

4.2.1 Experimental Design and Field Data

The experimental design and experimental procedure is described in Section 2.3 of Chapter 2. In this chapter, we are interested in how pupils' answering behavior changes if pupils can win a reward compared to pupils in the Control Treatment. Therefore, we distinguish between *incentivized* and *non-incentivized* pupils in our analysis. The group of *incentivized* pupils consists of the Fixed and Choice Treatment and the group of *non-incentivized* pupils are the pupils in the Control Treatment.[16] Thus in our further analysis we use a dummy variable indicating whether pupils belong to one of the incentivized treatments or to the Control Group.

Field data

To complement the results of our experiment, we received aggregate data on answers in the official *Känguru-Wettbewerb* test in 2013. For each test question, we know the absolute number of pupils that answered the question correctly and the number of pupils that omitted that particular question. We received this information by gender, school grades (from 3 to 13) and prize winners. Prize winners are the best 5% participants in each school grade. The field data include 780.085 pupils (approx. 7% of all pupils in Germany) in grades 3 - 13 of around 9.500 schools. A caveat of the data is that participation in the *Känguru-Wettbewerb* is voluntarily and that information on the school type is not available. The test is carried out simultaneously at the same day and time in all schools. Pupils in grades 3 - 6 have to solve 24 questions and pupils in grades 7 - 13 answer 30 questions. Furthermore, the majority of test takers attend High School. This information was given by the organizers of the *Känguru-Wettbewerb* in informal talks.

These data allow us to (a) compare the results of our baseline treatment with a broad sample of German pupils and to (b) test for gender differences in skipping test items over school grades.

4.3 Descriptive Statistics

4.3.1 Experiment

Our primary variable of interest is the number of skipped questions. Moreover, we compare gender difference in skipping between pupils in the incentivized and in the Control Group. Our identification of the average difference of skipped questions between boys and girls relies on our block randomization strategy. The most important control variable is pupils' last midterm grade. The last midterm grades are reported by teachers and available for almost all pupils. Midterm grades in Germany combine the written and verbal performance wherein the written part has a larger influence on the final grade; thus, these grades are a good measure of math ability. Importantly, the midterm grades can be treated as exogenous in our analysis because they were given to the pupils before teachers learned about the experiment.

[16] The interpretation of our results on skipped test items does not change if we do not group the incentivized treatments.

Additional control variables on pupil-level are gender, parents' education, an indicator variable on how much pupils like math and a dummy whether pupils are in grade 5 or 6. The latter variable controls for pupils' age and educational level at the same time. Parents' educational level is captured by the number of books at home (see Fuchs and Wößmann 2007; Wößmann 2005). The degree of how much pupils like math is self-reported and measured on a 1–5 scale. This measure approximates pupils' intrinsic motivation to answer questions. Pupils who like math should skip fewer questions because they have inter alia more fun in solving mathematical questions.

Table 2.1 in Section 2.4 compares the descriptive statistics to the actual data in North Rhine-Westphalia (NRW) and shows that our data are consistent with key school indicators from NRW. On average, subjects in our sample are 11.16 years old and have 0.92 older siblings. 43.49% of the subjects are female and 58.17% speak only German at home, while 37.59% speak another language and 4.24% speak two languages at home. The average midterm grade in mathematics is 2.86 on a scale from 1 to 6, where 1 is the highest and 6 is the lowest grade.

4.3.2 Field

The data of the official *Känguru-Wettbewerb* give us the opportunity to shed light on how gender differences in omitting test items evolve over school grades. To our knowledge, this is the first study using data on almost all grades within a school system—only the first two school grades are missing. Furthermore, we can compare our experimental results with a broad sample of German pupils.

Table 4.1 reports the number of pupils and proportion of prize winners that participated in the *Känguru-Wettbewerb* in 2013 separately for boys and girls.[17] There are more boys than girls taking the test in each grade and the overall number of participants is declining after grade 6. Column 3 in Table 4.1 is the share of boys who won a prize on all participating boys. Column 4 similarly reports on the share of female prize winners and Column 6 shows the proportion of male prize winner on all prize winners (male and female). The proportion of male prize winners over all males is constant over school grades but the proportion of female prize winners over all females is declining. This is unexpected as due to voluntary participation and hence self-selection the average ability of females should not be lower in higher grades. As there is a (low) cost of 2 Euro and participants in higher grades are often regular participants (writing the test each year)[18], the intrinsic motivation towards math should be higher for participants in higher grades.

[17] We dropped data of grade 13 as the number of participants is very low (only 258 observations all over Germany) and pupils in that grade usually prepare for their *Abitur* (High School graduation exam).

[18] This is anecdotal evidence given by the organizers of the *Känguru-Wettbewerb*.

Table 4.1: Descriptive Statistics - Field

	Number of Pupils		*Share of Winner*			
	Overall	Share of Girls	Boys [%]	Girls [%]	Difference [%]	Share Boys [%]
Grade						
3	106.528	0.4608	6.46	4.99	1.47	60.23
4	118.697	0.4749	6.18	4.81	1.37	58.69
5	154.772	0.4968	7.19	4.25	2.94	63.17
6	154.228	0.4909	6.69	4.38	2.31	61.32
7	94.078	0.4779	6.34	4.13	2.21	62.63
8	66.770	0.4630	6.67	3.88	2.79	66.60
9	45.422	0.4353	6.92	3.21	3.71	73.69
10	25.711	0.4123	6.62	3.10	3.52	75.23
11	10.677	0.3752	7.26	2.77	4.49	81.34
12	2.944	0.3601	7.59	1.98	5.61	87.20

Note: This table reports on the number of boys and girls who participated in the *Känguru-Wettbewerb* 2013 by school grade (columns 1-2). Columns 3-6 represent percentages and report on prize winners (top 5 %). Column 3 [4] represents the proportion of male [female] winners on the number of participating boys [girls] in the respective grade. Cell entries in column 5 shows the difference between column 3 and 4 and column 6 reports the proportion of male prize winner on the number of all prize winners (males + females).

4.4 Results

In this section, we first derive our hypotheses, we then focus on differences in answering test items between Vocational Schools and High Schools and present descriptive statistics on answering patterns separately for each difficulty section. Gender differences in skipping test items are studied thereafter and we estimate ordinary least square regression controlling for pupil characteristics. Next, we complement our experimental data with data of the nationwide test to investigate whether gender differences exist over all school grades. Finally, we examine the impact of skipping on test performance to evaluate whether different answering strategies result in different test outcomes. Thereafter, we discuss potential mechanisms which could explain our results.

4.4.1 Hypothesis

Answering strategies in multiple-choice tests have been analyzed with respect to gender differences in skipping items which has been attributed to attitudes towards competitive environments or risk-aversion. However, there is little *experimental* evidence on how skipping test items correlates with pupils' socio-economic background, although family background has been found to affect significantly educational outcomes and skills formation (see Anger and Schnitzlein 2016; Heckman et al. 2013 and the literature mentioned therein). In a recent paper, Almås et al. (2016) find that personal characteristics and family background are of great importance in explaining school dropout—having a parent with college education strongly reduces the likelihood of dropping out from the college track. Moreover, the link between differences in thinking styles and socio-economic background has been investigated in various fields of research (for a causal evidence see Mani et al. 2013). Cooper and Stewart (2013) show in a meta-analysis of 34 studies that cognitive ability differs with family income. When concentrating on experimental studies, they find that effect sizes associated with a US$ 1,000 increase in income ranged from 5% to 27% of a standard deviation for cognitive outcomes. Zhang and Postiglione (2001) find that those students who reported using thinking styles that are creativity generating and more complex and those who reported higher self-esteem tend to be students from families with higher socio-economic background.

There is also evidence in the literature that individuals who differ in cognitive ability vary in important personal traits which influence the skipping behavior in a test such as risk aversion and impatience (Dohmen et al. 2010).[19] In an experimental study among adults in Germany, Dohmen et al. (2010) find that people with lower cognitive ability are significantly more risk-averse and impatient. Deckers et al. (2015) show that children from families with higher socio-economic status are more patient, less likely to be risk-seeking, and score higher on IQ tests, while Sutter et al. (2013) can relate time preferences—a related concept to impulsive decision making—to field behavior.

As the literature indicates, cognitive ability and peer group composition is likely to shape individuals' personality traits and thus could effect their skipping behavior in a test. In the German school system, the transition recommendation to which school type to send a child is given by the elementary school and is based on talent and performance (i.e., grades), social skills and social behavior and motivation as well as learning virtues (Anders et al. 2010). Additionally, parents from a privileged background put more emphasis on sending their children to academically advanced school types than parents with lower socio-economic status. We therefore argue that the school type is a good—although not perfect—proxy of pupils' socio-economic background and their (cognitive) ability. Thus, if socio-economic groups differ in their cognitive and non-cognitive skills, they should also be likely to differ with respect to test taking strategies and skipping behavior in multiple-choice exams.

Hypothesis 1 *Pupils in Vocational School are more likely to give intuitive - automatic - answers whereas pupils in High School should have a preference to give accurate answers.*

Test taking strategies could not only differ by socio-economic background but also by gender. Turning to gender differences, Dohmen et al. (2010) find that the correlation between risk aversion and cognitive ability is—statistically not significant—smaller for women than for men. More importantly, economic studies have also shown that gender differences in personality traits could be shaped by nurture—the culture (Gneezy et al. 2009) or environment (Booth and Nolen 2012) of individuals. Booth and Nolen (2012) show that gender differences in behavior under uncertainty is influenced by peer group composition—same-sex or mixed-gender peer—and might reflect social learning rather than inherent gender traits. Among students in the UK from grades 10 and 11, the authors find that girls who attend single-sex schools were more likely to take risk. These studies suggest that there are

[19] Furthermore, Borghans et al. (2009) show that psychological traits are strongly associated with risk but not with ambiguity.

gender differences in answering risky test questions which could result in girls being more likely to skip test items. Baldiga (2014) and Pekkarinen (2015) document that girls perform worse than boys by answering too few multiple-choice questions and i.e. that girls therefore are less successful in (multiple-choice) university entrance exams. However, little is known about the correlation of the school type and the gender gap in the willingness to guess, although recent studies suggest that the gender gap in school varies by family background. Autor et al. (2016) find that family disadvantage disproportionately negatively affects the academic and behavioral outcomes of school-age boys relative to girls but that better-quality schools help to mitigate this gender gap. We therefore expect the gender gap in skipping to be larger for pupils in Vocational School relative to pupils in High School.

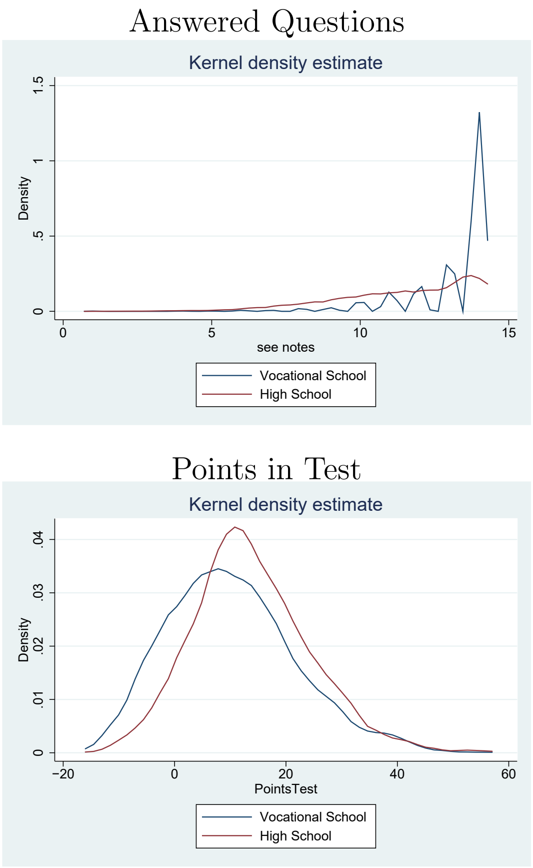
Hypothesis 2 *Girls skip more test items than boys but this gender gap is larger in Vocational Schools than in High Schools.*

4.4.2 Differences in Skipping Test Items by Type of School?

We analyze whether pupils in different school types, this means pupils with an on average different socio-economic background, apply different answering strategies in a multiple-choice test. As pupils in Vocational School and High School differ in their initial ability, we designed one test for pupils in Vocational School and one test in High School. As discussed later, we do not find evidence, that the test for pupils in High School was disproportionately more difficult than the test in Vocational School.

Figure 4.1 presents kernel density estimates for the number of answered test items and the number of points in the test for Vocational School and High School. While pupils in Vocational School tend to give more answers, pupils in High School get higher test scores. Testing the equality of score means shows that differences between school types are significant for both, the number of answered questions ($p < 0.001$) and points in the test ($p < 0.001$).

Figure 4.1: Number of Answered Questions and Points Received in Test



Note: Figure (a) presents kernel density estimates for the number of answered test questions for High Schools and Vocational Schools. Figure (b) presents kernel density estimates for the number of points gained in the test for High Schools and Vocational Schools.

Table 4.2 gives a more detailed picture on the answering strategies of pupils in High School and Vocational School and reports on the average percentage of skipped items, the share of correctly answered questions and the amount of points pupils received in the test. Across difficult sections, pupils in Vocational School omit roughly only one third as many test items as pupils in High School. Pupils in Vocational School answer more questions but at the same time the share of correctly answered questions is lower than for pupils in High School. One explanation for the difference in skipping items between school types could be that the test in High Schools is more difficult than the test in Vocational Schools and thus pupils in High School have not the time to answer all questions. This can, however, be checked by investigating how many answers pupils skipped in the last section of the test. In the difficult section, we see that only 1.36% of pupils in High School did not answer a single question and only 10.19% did not answer the last two questions. This is clearly an indication that pupils in High School had enough time to answer questions and that the difference in skipping between High Schools and Vocational Schools is due to other reasons which are discussed in Section 4.5. To summarize, pupils in High School skip more questions but at the same time the share of correctly answered questions and overall performance is higher than in Vocational School. This suggests that Hypothesis 1 can be confirmed.

Result 1 *Answering strategies differ by school type. While pupils in Vocational School tend to answer almost every question, pupils in High School seem to answer question more accurately.*

4.4.3 Gender Differences in Skipping Test Items and Test Performance

We now turn to *gender* differences in multiple-choice tests. We first present results on gender differences in skipping test items for our experimental data and examine gender difference for all school grades using field data of the official "*Känguru-Wettbewerb*". We then relate differences in skipping items to test performance.

Experiment The dependent variable of the analysis is the number of skipped questions. To estimate gender differences in the willingness to skip test items we estimate a linear model (OLS). We cluster the standard errors on classroom level—the level of randomization—and use interaction terms between pupils' gender and whether pupils are incentivized. As OLS might not be efficient due to the data being non-negative and overdispersion in the High School sample (Vocational School: $\ln\alpha = -0.120$, p-value = 0.579; High School: $\ln\alpha = -0.496$, p-value = 0.001)[20], we provide robustness checks using a negative binomial specification in Appendix 4.7. As the results change neither in significance nor in size and OLS coefficients make a straightforward interpretation, we decided to report those. Furthermore, we estimate the models separately for High Schools and Vocational Schools linking them by seemingly unrelated estimations and allow for school fixed effects. We use seemingly unrelated estimation as it combines the parameter estimates, the variance and covariate variances of the separately estimated equations into a robust single parameter-vector and simultaneous variance-covariance matrix. The advantage of seemingly unrelated estimations is the robustness to cross-equation correlation and between group heteroskedasticity; consequently, it can overcome the problem of multiple testing. We estimate the following linear model:

$$\begin{aligned} Omittedi = \;& \beta_0 + \beta_1\ Incentivized_i + \beta_2\ Female_i + \beta_3\ Incentivized_i\ X\ Female_i \\ & +\gamma\ Covariates_i + \delta School_i + \varepsilon_i \end{aligned} \tag{4.2}$$

$Omitted_i$ is the number of omitted questions in the test by pupil i, $Incentivized_i$ is a dummy indicating whether pupils are in a treatment group, $Female_i$ indicates whether pupil i is female, $Covariates_i$ is a vector of characteristics of pupil i: the midterm grade, number of books at home, whether pupil i is in grade 5 or 6 and

[20] The distribution of the number of omitted questions is presented in Figure 4.3 in Appendix 4.7.

Table 4.2: Average Number of Omitted Questions, Probability of Success and Points achieved by Difficult Levels and by Incentivized

	Omitted questions (in percent)				Probability of success (in percent)				Aver. Points per Question			
	Easy	*Medium*	*Difficult*	*Overall*	*Easy*	*Medium*	*Difficult*	*Overall*	*Easy*	*Medium*	*Difficult*	*Overall*
Not Incentivized												
Vocational School	3.39	11.20	8.74	6.59	38.18	23.41	47.08	39.09	0.49	0.15	1.66	0.68
High School	13.06	22.98	26.45	20.43	62.05	46.57	40.26	50.94	1.28	0.98	0.90	1.06
Incentivized												
Vocational School	3.80	11.00	7.12	6.13	38.42	28.19	52.29	39.23	0.51	0.35	1.94	0.79
High School	10.05	16.66	20.40	15.37	60.62	37.90	36.30	46.14	1.27	0.74	0.84	0.96

Note: This table reports the percentage number of skipped question (*Omitted questions*), the probability to answer a question correct—the share of correct answers on all given answers—(*Probability of success*) and the average number of points per question (*Aver. Points per Question*) for each section differentiated by school type and incentive groups.

an indicator how much pupil i likes math, $school_i$ controls for school fixed effects and ε_i is a stochastic i.i.d. error term. We include school fixed effects to control for unobserved school specific effects.

Table 4.6 in Appendix 4.7 reports basic descriptive statistics of omitted questions over all difficulty sections. The average number of omitted questions varies between school types and for incentivized and non-incentivized pupils. Additionally, Figures 4.4 and 4.5 in Appendix 4.7 show that there is a noticeable difference in omitting questions in High Schools between girls and boys and between incentivized and non-incentivized pupils. Further evidence that differences in skipping items occur only in High Schools is given in Table 4.6 (Panel B) in Appendix 4.7. A Mann-Whitney test allows us to reject the null hypothesis that the samples are drawn from the same distribution between all tested pairs in High School. Females always tend to skip more questions than males no matter if they are incentivized or not. However, incentivized pupils tend to skip fewer questions than non-incentivized pupils.

Table 4.3 presents estimates of the gender gap which is captured by *Female* and is the difference between the number of questions omitted by girls and the number of questions omitted by boys.[21] Panel A shows the average effect of gender on the number of omitted questions, the effect of rewarding pupils (*Incentivized*) and the interaction effect of these two covariates. Panel B summarizes the answering gap distinguished by incentivized and non-incentivized pupils.

Overall, pooling difficulty levels (Q1-Q14), we find that girls in High School skip significantly more questions than boys if they are non-incentivized (0.922, $p < 0.001$). In contrast the answering gap is closed if pupils are incentivized (0.239, $p = 0.229$). There is no significant difference in Vocational School for incentivized (-0.117, $p = 0.101$) and non-incentivized pupils (0.101, $p = 0.299$). Splitting the sample by difficulty sections, we do not find a gender gap in the easy section for incentivized and non-incentivized pupils in High School and Vocational School. The gender gap in High School occurs only for non-incentivized pupils at the 10% level in the medium section (0.295, $p = 0.094$) and seems to be strongest in the difficult section (0.553, $p < 0.001$). Surprisingly, an answering gap occurs although skipping becomes less attractive in the difficult section compared to the easy section. One reading of this result is that girls may become underconfident due to a stereotype-threat—as the difficulty level of sections was made salient at the beginning of each section. However, we cannot rule out that boys are simply overconfident.

Result 2 *Although attractiveness of guessing is increased for difficult questions, girls in High School skip more questions than boys. This gap can be closed by providing extrinsic incentives for performance. Moreover, there is no gender gap for easy questions and for pupils in Vocational School.*

Field We now analyze whether a gender gap exists across school grades using data from the official *Känguru-Wettbewerb 2013.* These data are comparable to our

[21] Table 4.7 in Appendix 4.7 presents raw treatment effects of OLS regression separately for each section.

Table 4.3: Number of Omitted Questions

	Overall (Q1-Q14)		Easy (Q1-Q5)		Medium (Q6-Q10)		Difficult (Q11-Q14)	
Panel A: Regression	*Vocational School*	*High School*	*Vocational School*	*High School*	*Vocational School*	*High School*	*Vocational School*	*High School*
Treatments								
Incentivized	0.110	-0.488**	0.032	-0.141*	0.071	-0.264**	0.007	-0.083
	[0.111]	[0.244]	[0.043]	[0.079]	[0.048]	[0.110]	[0.040]	[0.095]
Female	0.101	0.922***	0.021	0.074	0.075	0.295*	0.004	0.553***
	[0.097]	[0.281]	[0.049]	[0.074]	[0.049]	[0.176]	[0.030]	[0.099]
Female × Incentive	-0.218*	-0.684**	-0.058	-0.011	-0.122**	-0.226	-0.037	-0.446***
	[0.120]	[0.353]	[0.058]	[0.095]	[0.058]	[0.202]	[0.040]	[0.135]
Controls								
SchoolFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1193	867	1193	867	1193	867	1193	867
Panel B: Contrasts	*Gender Gap*							
Not Incentivized	0.101	0.922***	0.021	0.074	0.075	0.295*	0.004	0.553***
	[0.097]	[0.281]	[0.049]	[0.074]	[0.049]	[0.176]	[0.030]	[0.099]
Incentivized	-0.117	0.239	-0.037	0.063	-0.047*	0.069	-0.033	0.107
	[0.072]	[0.199]	[0.034]	[0.069]	[0.028]	[0.084]	[0.028]	[0.087]

Note: Panel A reports the results of least squares regressions separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. The gender gap in skipping questions is captured by *Female* (Female=0: boys; Female=1: girls). Panel B reports the gender difference in skipping questions for incentivized and non-incentivized pupils resulting from Panel A. Dependent variable: number of skipped questions. Covariates: last midterm grade, number of books at home, math curiosity (self-reported on 1-5 scale), academic year (grade 5 or 6). Robust standard errors are reported in parentheses and clustered on classroom-level. As OLS is not efficient due to overdispersion of our data, we provide robustness checks using a negative binomial specification provided in Appendix 4.7. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

(non-incentivized) baseline treatment as the test does not count towards the math grade and hence is low-stakes testing. However, in the *Känguru-Wettbewerb* the top 5% of all participating pupils within a grade receive a small prize. Nevertheless, this setting unlikely increases the stakes of the test for three reasons: First, pupils are competing against all pupils of the same school grade in Germany. Hence, the perceived likelihood of success in the competition should be low. Second, prizes are "paid" with a delay of about 1-2 month and Levitt et al. (2016) show that all motivating power of incentives vanishes when rewards are handed out with a large delay. Third, it is unlikely that pupils participate with the aim of winning these prizes because they consist i.e. of experiment kits, interesting strategic games, challenging mathematical puzzles and books—selected on the basis that they are mentally challenging, stimulating and appropriate to promote creativity as well as social behavior. There is evidence that these kind of incentives—*Mastery Goal Incentives*—are unpopular among pupils in Germany (see Chapter 2).

Table 4.4 summarizes the share of omitted questions separately for boys and girls by school grade and separately by degree of difficulty. For each question in the test we received data on how many pupils answered the question correctly or omitted the answer and calculated the shares for each respective test question. The cell entries in Table 4.4 represent the mean over all shares in the respective test and the number of test takers is reported in parentheses. The significance of differences between boys and girls is estimated by testing on the equality of proportions. Overall, we see that boys as well as girls in higher grades tend to skip more questions than boys and girls in lower grades and that the gender gap tends to increase by approx. 3% from grade 3 to grade 12. This increase seems to be driven mainly by the medium and difficult section. We also observe that the gender gap in the difficult section is higher compared to the easy section in almost every grade with the exception of grades 7 and 8. These findings are in line with the findings of our experiment. Moreover, we find that the accuracy level—correctly answered questions—does not seem to decrease by school grade (see Table 4.9 in Appendix 4.7 for gender differences in correct answers by school grade). The share of correct answers is roughly constant for girls but the accuracy level of males tends to increase by about 4% which consequently increases the gender gap in correct answers.

The field data allow us to differentiate pupils by ability—prize winner and non-prize winner—and to answer the question whether the gender gap persists among top performers? Table 4.10 in Appendix 4.7 summarizes the share of omitted questions and correct answers for prize winners (top 5%) and non-prize winner for all school grades. We see that the gender gap seems to be driven solely by non-prize winners. In contrast, the gender gap in skipping answers tends to be (not significantly) reversed for top performing pupils in higher grades. In lower grades—grades 3 to 6—prize winning girls significantly skip more questions than prize winning boys (approx. 1%) but in higher grades there is no significant gender gap. However, the gender gap coefficient is negative in grade 12, indicating that that girls answer more questions than boys. Nevertheless, differences in the number of correct answers range from -0.90 to -1.65 for top performers but are statistical not significant.

To summarize, boys as well as girls tend to skip more answers in higher grades than boys and girls in lower grades. Moreover, the difference between boys and girls in skipping test items and in answering correctly tends to increase over school grades. These results seem to be driven solely by non-prize winners. However, limitations of our field data is that results could be driven by self-selection due to voluntarily participating in the test. As the number of participants is decreasing over school grades it is most likely that highly motivated pupils keep on taking the test in higher grades.

Result 3 *The gender gap in skipping test items exists in all school grades for non-prize winner but not for the very top performers.*

Impact of guessing on test performance

The SAT scoring rule has been redesigned recently to a "rights-only" scoring method to encourage students to guess and Akyol et al. (2016) find that if there is no negative marking in the Turkish University Entrance Exam—so that guessing when in doubt is optimal—increases women's representation in the top 5% of placements by 0.3%. Furthermore, Pekkarinen (2015) shows that girls are less likely to gain entry in Finnish university because they omit too many items in the entrance exam and Baldiga (2014) shows for practice questions from SAT II history test that because girls skip more questions, they receive lower test scores than boys with the same knowledge of the material. In our experiment, girls in High School skip significantly more test items than boys if questions are difficult. As the previous literature points out, this differences in test answering strategies need to be discussed within the context of performance as a student usually chooses a test strategy to maximize her performance. Hence, if girls would be more efficient in skipping questions than boys, it might be that girls and boys employ different answering strategies but arrive at similar outcomes. We therefore exploit whether a decline in *performance* could be explained by omitting more questions.

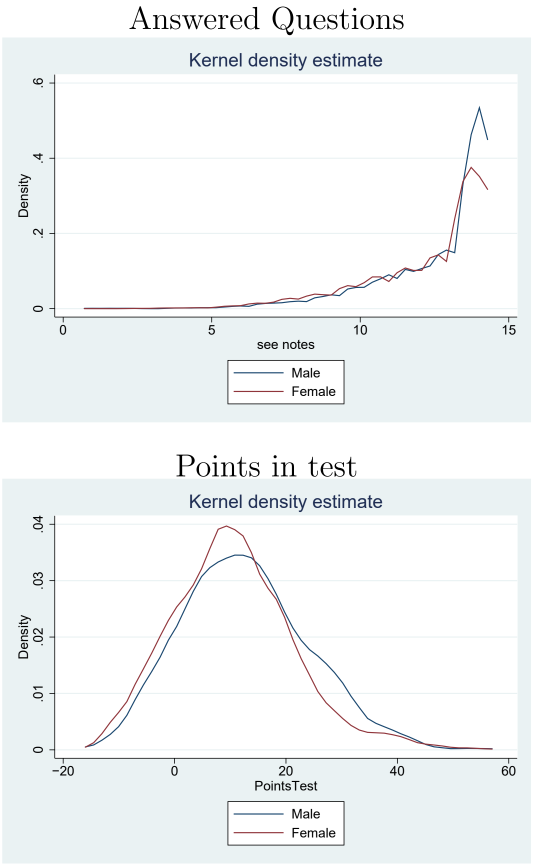
Figure 4.2 suggests a negative correlation between skipping test items and performance in our sample. The kernel density estimates on the number of answered questions and points in test for boys and girls indicate that boys tend to answer, on average, more questions than girls and score higher in the test.

Table 4.4: Omitted Answers by Academic Year

Dep. Var:	*Overall*			*Easy Section*			*Medium Section*			*Difficult Section*		
Omitted Answers	Boys	Girls	Difference	Boys	Girls	Difference	Boys	Girls	Difference	Boys	Girls	Difference
Grade												
3	15.77	16.78	1.01***	6.15	6.69	0.54***	12.30	13.10	0.80***	28.86	30.56	1.70***
	(57440)	(49088)		(57440)	(49088)		(57440)	(49088)		(57440)	(49088)	
4	10.68	11.30	0.62***	3.15	3.49	0.34***	6.54	6.83	0.29**	22.35	23.59	1.24***
	(62324)	(56373)		(62324)	(56373)		(62324)	(56373)		(62324)	(56373)	
5	14.82	15.66	0.84***	5.53	6.33	0.80***	13.83	14.74	0.91***	25.11	25.91	0.80***
	(77888)	(76884)		(77888)	(76884)		(77888)	(76884)		(77888)	(76884)	
6	11.05	11.19	0.14	3.54	3.73	0.19**	9.69	9.69	0.00	19.91	20.16	0.25
	(78514)	(75714)		(78514)	(75714)		(78514)	(75714)		(78514)	(75714)	
7	20.51	21.62	1.11***	10.56	11.97	1.41***	20.80	22.23	1.43***	30.18	30.65	0.47
	(49118)	(44960)		(49118)	(44960)		(49118)	(44960)		(49118)	(44960)	
8	17.28	18.98	1.70***	7.38	9.19	1.81***	17.02	19.25	2.23***	27.43	28.50	1.07***
	(35858)	(30912)		(35858)	(30912)		(35858)	(30912)		(35858)	(30912)	
9	21.36	23.42	2.06***	7.14	8.93	1.79***	22.62	24.33	1.71***	34.32	37.00	2.68***
	(25649)	(19773)		(25649)	(19773)		(25649)	(19773)		(25649)	(19773)	
10	21.02	23.11	2.09***	6.37	8.33	1.96***	21.47	23.37	1.90***	35.23	37.64	2.41***
	(15110)	(10601)		(15110)	(10601)		(15110)	(10601)		(15110)	(10601)	
11	23.16	26.27	3.11***	6.66	8.70	2.04***	23.35	27.28	3.93***	39.47	42.84	3.37***
	(6671)	(4006)		(6671)	(4006)		(6671)	(4006)		(6671)	(4006)	
12	24.73	29.13	4.40***	7.14	9.96	2.82***	25.08	30.48	5.40***	41.98	46.96	4.98***
	(1884)	(1060)		(1884)	(1060)		(1884)	(1060)		(1884)	(1060)	

Note: This table reports the averages of omitted answers in the *Känguru-Wettbewerb* test 2013. Cell entries report percentages and the number of pupils that participated in the test is reported in parentheses. *Difference* is the difference girls − boys. The significance of gender differences is estimated by testing on the equality of proportions.

Figure 4.2: Number of Answered Questions and Points Received in Test by Gender


Answered Questions

Points in test

Note: Figure (a) presents kernel density estimates for the number of answered test questions for males and females. Figure (b) presents kernel density estimates for the number of points gained in the test for males and females.

Table 4.5 presents the OLS estimates on the effect of the share of omitted questions on the number of points for each difficulty section. We control *inter alia* for pupils' gender, whether pupils are incentivized and for school fixed effects. The variable of interest is *Share Omitted* as it shows the impact of skipping more test items on performance. Overall (Question 1 - Question 14), we find that omitting more questions decreases significantly test performance in High School (-6.018, p = 0.001) but not in Vocational School (-1.347, p = 0.411). This detrimental effect of skipping is significant for all three difficult sections in High School and also for the difficulty section in Vocational School. Furthermore, holding all other variables constant, we find that girls perform poorer than boys in High School if questions are difficult (*Female*: -1.110, p = 0.043) but not if questions are of low or medium difficulty. One explanation could be that girls suffer from a stereotype-threat because coefficients are positive and insignificant in the easy and medium section. However, we cannot link causally the poorer performance of High School girls in the difficult section to a stereotype-threat. Nevertheless, as the interaction term of female and incentivized is positive and significant, the performance gap can be closed if girls are rewarded for performance. This result suggest that a stereotype-threat is indeed a good candidate to explain performance differences as the reward may shift the focus of item difficulty to winning the reward. However, we discuss other explanations in the following section.

Further (descriptive) indication that skipping difficult questions is detrimental for girls performance in High School can be found in Table 4.8 in Appendix 4.7. Non-incentivized High School girls omit 34.09% of all difficult questions whereas incentivized girls skip only 22.09%. On the other side, non-incentivized girls get on average 0.58 points per difficult question but incentivized girls receive on average 0.84 points per difficult question.[22]

Result 4 *Skipping more test items decreases test performance for pupils in High School.*

4.5 Discussion

Our data suggest that pupils in different school types apply different answering strategies and that girls in High School skip more multiple-choice questions than boys. This gap occurs mainly if questions are difficult. In the following we first discuss why pupils of different school types might apply different answering strategies and then discuss potential causes for the gender gap in skipping.

Answering strategies in school types

While pupils in Vocational School answer more questions than pupils in High School, they obtain lower test scores. One reading of this answering pattern could be that pupils from low income families differ in their way of thinking compared to pupils from high income families. Following the terminology used by Kahneman (2003), pupils in Vocational School seem to answer questions accordingly to the automatic System 1 while pupils in High School seem to answer according to the effortful System 2.[23] In a similar vein, Heller et al. (2016) argue that automaticity interacts with the social environment. The authors hypothesize that in poor neighborhoods there is more variability in the type of automatic response that is adaptive to the different situations people encounter (e.g. “street life” situation and “school life” situation) and that these automatic responses might be adaptive in some but not all situations.

[22] Non-incentivized boys and incentivized boys do not differ significantly in terms of skipping difficult questions (20.08% vs. 18.95%) but non-incentivized boys get on average more points in the difficult section(1.19 vs. 0.85).

[23] The distinction between two types of cognitive processes—the ways the brain forms thoughts—and the labels of System 1 and System 2 have been emphasized by many researchers in psychology (Chaiken and Trope 1999; Epstein 1994; Kahneman and Frederick 2002). System 2 is the conscious and reasoning self, performing effortful mental activities including complex computations. System 1 is the brain’s fast, automatic and intuitive approach. While System 2 is the mind’s slower, analytical mode, where reasoning dominates, System 1 is more influential, guiding System 2. Hence, in effortful tasks like a mathematical multiple-choice test, System 2 is more likely in making better decisions than System 1. The concept of System 1 and System 2 thinking has also received recognition by economists (Shleifer 2012; Lavecchia et al. 2016).

Table 4.5: Impact of Skipping on Test Performance

	Overall (Q1-Q14)		Easy (Q1-Q5)		Medium (Q6-Q10)		Difficult (Q11-Q14)	
Dep. Var: Points in Test	*Vocational School*	*High School*	*Vocational School*	*High School*	*Vocational School*	*High School*	*Vocational School*	*High School*
Treatments								
Share Omitted	-1.347	-6.018***	0.566	-3.594***	-0.264	-3.449***	-3.231***	-1.621***
	[1.639]	[1.772]	[1.010]	[0.957]	[0.282]	[0.968]	[0.467]	[0.586]
Incentivized	1.342**	-0.707	0.115	0.401	0.611***	-0.777*	0.617**	-0.439
	[0.558]	[1.001]	[0.398]	[0.449]	[0.227]	[0.469]	[0.264]	[0.445]
Female	0.265	-0.490	0.228	0.257	-0.093	0.448	0.130	-1.110**
	[0.519]	[1.144]	[0.492]	[0.765]	[0.190]	[0.545]	[0.364]	[0.548]
Female × Incentive	-1.149	-0.189	-0.348	-1.063	-0.409	-0.417	-0.429	1.410**
	[0.716]	[1.389]	[0.585]	[0.850]	[0.306]	[0.618]	[0.469]	[0.660]
Controls								
SchoolFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	1193	867	1193	867	1193	867	1193	867

Note: This table reports the results of least squares regressions separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. Dependent variable: number of points in the test. Covariates: last midterm grade, number of books at home, math curiosity (self-reported on 1-5 scale), academic year (grade 5 or 6). *Share Omitted* is the share of skipped answers and measures the impact of skipping questions on performance in the test. Robust standard errors are reported in parentheses and clustered on classroom-level. * p<0.10, ** p<0.05, *** p<0.01

In an educational context, Lavecchia et al. (2016) argue that the framework of System 1 and System 2 thinking can explain why pupils may invest too little in their education today. System 1 evaluates feelings today whereas System 2 anticipates feelings in the future. Behavior that is biased due to System 1 can have important implications for education. When faced with the decision of doing homework for an extra hour or enjoy time with friends, pupils whose decisions are driven by System 1 may decide for the latter option as they are likely to overemphasize the costs of studying relative to the potential future benefits (Lavecchia et al. 2016). The role of automaticity can also be transferred to multiple-choice testing. Pupils who answer according to System 1 could come up quickly and intuitively with an answer whereas pupils of System 2 thinking act more sophisticated and review the proposed answer by System 1. Pupils in High School omit more questions but have a higher share of correctly answered questions. In contrast, pupils in Vocational School skip only few questions but have a lower share of correctly answered questions. This indicates that—in terms of the two system thinking—pupils in Vocational School are guided by System 1 and therefore avoid effortful thinking and answer according to effortless intuition and that pupils in High School are guided by a more sophisticated System 2 thinking.

A further candidate to explain the data is self-esteem. Pupils in High School could have a greater disutility of answering questions incorrectly due to a higher degree of self-esteem which in turn would result in a higher share of correctly answered questions. In a meta-analytic review of 446 samples, Twenge and Campbell (2002) show a positive relationship between socio-economic status and self-esteem. Individuals with higher socio-economic status scored higher on measures of self-esteem but expressed in terms of a correlation, the weighted effect was small (0.08).

Socio-economic status seems to have an effect on self-esteem and thus the way of answering questions in a multiple-choice test. However, effect sizes found by Twenge and Campbell (2002) were only small. Moreover, gifted students seem to be less likely to overestimate their math performance and overestimate it to a smaller degree than regular math students (Pajares and Graham 1999). This could also explain the fact, that pupils in Vocational School answer almost every question but have a lower share of correctly answered questions than pupils in High School.

Gender gap due to higher-stakes, risk aversion or overconfidence?

In the following, we discuss whether risk aversion, overconfidence or different responses to the stakes of the testing environment could explain *gender* differences in guessing in High Schools.

Higher-Stakes Acceptance to university is based to a large extent on the (multiple-choice) university entrance exam and the number of accepted applicants is limited. This setting creates a high-stakes testing environment and females may perform worse as they dislike high pressure situations or competitive settings especially in mathematical tasks (Niederle and Vesterlund 2010; 2007). In our experiment, we exogenously vary the stakes of the test and examine whether the answering gap in High Schools increases in stakes. In our control treatment stakes of the test were reduced to a minimum, as the test did not enter the main grade. We then increased the stakes by incentivizing pupils with an external reward for higher performance.[24] As we do not find an answering gap in the treatment group, it is unlikely that choking under pressure due to the higher stakes of the testing environment is a driver of the answering gap—in this case as gender differences should be higher for pupils in the incentivized groups.

Risk Aversion Gender differences in risk taking have been frequently mentioned to explain the answering gap in multiple-choice tests. Answering questions without surely knowing the answer is a risky decision if points are deducted for incorrect answers. In varying the degree of punishment for skipped questions previous research has shown that some part of the answering gap can be explained by girls being more risk-averse than boys (see Tannenbaum 2012; Baldiga 2014). To analyze whether risk aversion may explain some part of the answering gap in our experiment, we focus on non-incentivized girls in High Schools.[25] Our test is designed such that omitting difficult questions is more attractive because we compensate risk-averse and less confident pupils for the increase in item difficulty by an increase of points for each correctly answered question. Moreover, we keep the punishment for incorrect answers constant (deducting always one point irrespective of item difficulty). Hence, a risk-averse subject should—at least—skip *not* significantly more questions in the difficult section than in the easy section. Table 4.8 in Appendix 4.7 presents the percentage number of skipped questions, the share of correctly answered questions and the number of points achieved in the test distinguished by pupils' gender, school type and treatment group. Considering the share of omitted questions of non-incentivized girls in High School—13.94% in the easy section, 27.07% in the middle section and 34.09% in the difficult section—we see that more questions are skipped with increased difficulty level. This in fact is an indication that girls might be risk-averse. However, a risk-averse subject would only answer a question if the probability to answer the question correctly is sufficiently high. This in turn should result into a higher share of correctly answered questions. Thus, if girls are indeed risk-averse, the share of correctly answered questions should—at least—not decrease in difficulty. However, the share of correctly answered questions of non-incentivized girls in High School is decreasing from 60.54% in the easy section to 45.15% in the middle section and to 36.53% in the difficult section. This suggests that girls in our sample seem to be not sufficiently risk-averse to explain the answering gap.[26]

Overconfidence Gender differences in overconfidence can result in a gender gap in skipping test items in two ways. First, girls may be underconfident—due

[24] Although stakes are increased, the incentivized test is still low stakes as the literature usually refers to high-stakes test if the test outcome has consequences for the final course grade.

[25] We focus on non-incentivized pupils in High School as this is the only school type in which we observed an answering gap.

[26] We find similar results for incentivized girls and—incentivized as well as non-incentivized—boys in High School.

to a stereotype-threat—which may cause girls to shy away from challenging tasks.[27] Second, boys may be overconfident. We find suggestive evidence that the answering gap in High Schools for non-incentivized pupils is driven by a stereotype-threat explanation as we observe the answering gap only for difficult questions and not for the easy section. Moreover, the answering gap vanishes once pupils are incentivized which could be due to: (i) boys skipping relatively more questions when incentivized or (ii) girls answering relatively more questions when incentivized. Examining the share of omitted questions (Table 4.8 in Appendix 4.7), we find evidence for the latter. Girls in High School skip significantly less difficult answers when incentivized (22.09% vs. 34.09%, $p < 0.001$) whereas the share of omitted answers does not significantly differ between incentivized and non-incentivized boys (20.80% vs. 18.95%, $p < 0.473$). In other words, incentivizing girls with an extrinsic reward shifts their focus from the level of difficulty to winning the reward. The stereotype-threat explanation would be also in line with the findings of the lab experiment by Coffman (2014). The author finds that individuals are less willing to contribute ideas—and hence knowledge—in areas that are stereotypically outside of their gender's domain. This is not driven by ability but is largely driven by self-assessments, rather than fear of discrimination. The findings by Coffman (2014) and the fact that our math test is stereotypically outside girls' domain is a further indication that our results could be driven by a stereotype-threat. We can also exclude that the answering gap occurs because difficult questions are just harder to answer for girls than for boys. In Table 4.8 in Appendix 4.7, we see that for incentivized pupils in High School the share of correctly answered questions in the difficult sections is higher for girls (36.06%) than for boys (35.76%). Gender differences in overconfidence are highly task dependent and men are more overconfident most strongly in masculine tasks. Furthermore, overconfidence is greatest for difficult tasks and tasks lacking fast and clear feedback (see the literature mentioned in Barber and Odean 2001 for a review on overconfidence). Thus, as the answering gap in our multiple-choice test occurs only for difficult tasks where girls did not get immediate feedback, overconfidence might indeed cause gender differences. Unfortunately, we can not directly identify if overconfidence is a driver of the answering gap in our experiment because we do not measure pupils' stated beliefs about their test performance.

[27] There is evidence that women self-select less often into demanding tasks compared to men (see Niederle 2016 for a literature review). In a laboratory experiment, Niederle and Yestrumskas (2008) analyze whether men and women of the same ability differ in their decision to seek challenges. Participants had to solve mazes of two difficulty level (hard and easy) for 10 minutes and were paid according to their performance. The experiment consisted of three rounds, where all participants were solving easy mazes in the first round. Based on task 1 performance participants were divided into two groups (i) those who have higher expected earnings from the easy task and (ii) those who have higher expected earnings from the hard task. In one treatment participants than had to choose the difficulty level of mazes for the following two rounds. Niederle and Yestrumskas (2008) find that there are no gender differences in performance, or beliefs about relative performance but that men choose the hard task about 50% more frequently than women, independent of performance level.

To summarize, we find an answering gap between boys and girls in High Schools which can be caused through multiple mechanisms. We find suggestive evidence that a large part of this gap could be explained by a stereotype-threat. However, we do *not* claim that this is the only explanation.

4.6 Conclusion

Using experimental data from a multiple-choice test in different school types in Germany, we analyze answering strategies of pupils from Vocational School and High School along with differences between boys and girls. Complementing the experimental data with aggregate data of a nationwide test, we also shed light on gender differences in skipping for pupils in all school grades.

Our results disclose structural biases in multiple-choice tests. First, pupils in Vocational School answer more questions than pupils in High School but perform lower in the test. Second, the multiple-choice testing format introduces a gender gap in skipping test items in High Schools if questions are difficult but no gender gap is found in Vocational Schools. Third, the gender gap in skipping test items persists over school grades. These findings confirm with Hypothesis 1 that pupils in Vocational School apply different answering strategies than pupils in High School. However, we find no support for Hypothesis 2 that the gender gap is larger in Vocational Schools compared to High Schools. Moreover, omitting test items decreases test performance in High Schools. We find suggestive evidence that the gender gap is due to a stereotype-threat. Girls in High School skip significantly more questions only if questions are difficult although the attractiveness of guessing is higher for difficult questions than for easy questions. Further support for a stereotype-threat explanation is the fact that the gender gap is closed if the difficulty of the task is made less salient by shifting the focus of pupils to winning an extrinsic reward. Nonetheless, we do *not* claim that a stereotype-threat is the only explanation for our results.

Conducting our experiment in Vocational as well as High Schools allows us to analyze the answering behavior in multiple-choice test for a sample which mirrors more closely the general population, while previous studies have mainly focused on high performing students or pupils who self-select into taking university entrance exams.

Recent studies have shown structural biases in multiple-choice tests and the College Board has already adjusted the SAT scoring rules to this biases. Our results confirm the existence of gender biases in multiple-choice testing. However, what has been missing in the literature so far is whether there is a school type bias. Our analysis suggests that pupils of different school types apply different answering strategies in multiple-choice testing. Therefore, educators and policy makers have to take this differences in school types into account to increase equality of promotion within the educational system and educational opportunities. Nevertheless, any testing format is advantageous for some types and less advantageous for others (oral exams may favor extroverts; open-ended questions may favor girls). It therefore needs future research on the optimal design of testing formats to reduce or close testing inequalities in education.

4.7 Appendix

Table 4.6: Descriptive Statistics on Skipped Answers

Panel A:	*Num. of Obs.*	*Mean*	*Share*	*Std. Dev.*	*Min.*	*Max.*
			Overall			
High School						
Incentivized	641	2.151	0.1537	2.179	0	10
not-Incentivized	242	2.860	0.2043	2.618	0	12
Vocational School						
Incentivized	864	0.552	0.0613	0.995	0	8
not-Incentivized	366	0.593	0.0659	1.026	0	5
			Easy Section			
High School						
Incentivized	641	0.502	0.1005	0.625	0	3
not-Incentivized	242	0.653	0.1306	0.754	0	3
Vocational School						
Incentivized	864	0.190	0.0380	0.460	0	4
not-Incentivized	366	0.194	0.0388	0.460	0	3
			Middle Section			
High School						
Incentivized	641	0.833	0.2083	1.027	0	5
not-Incentivized	242	1.149	0.2872	1.247	0	5
Vocational School						
Incentivized	864	0.220	0.1100	0.462	0	2
not-Incentivized	366	0.224	0.1120	0.449	0	2
			Difficult Section			
High School						
Incentivized	641	0.816	0.2040	1.006	0	4
not-Incentivized	242	1.058	0.2645	1.114	0	4
Vocational School						
Incentivized	864	0.142	0.0712	0.393	0	2
not-Incentivized	366	0.175	0.0874	0.421	0	2

Panel B: *Mann-Whitney test of independence*	*Vocational School*		*High School*	
Male vs. Female	0.188	(0.851)	-3.232	(0.001)
not-Incentiv. vs. Incentiv.	0.457	(0.648)	3.458	(0.001)
Male not-Incentiv. vs. Female not-Incentiv.	-1.949	(0.051)	-2.787	(0.005)
Male Incentiv. vs. Female Incentiv.	1.531	(0.126)	-2.168	(0.030)

Note: Panel A reports on absolute number, mean, standard deviations, minimum and maximum number of skipped questions separately by difficult section, school type and treatment groups. *Share* is the share of the number of omitted answers on the number of questions per section. As discussed in Section 4.2 the medium and difficult section in Vocational Schools consisted of 5 respectively 4 questions but in our analysis we included only the questions which belong originally to the respective difficult sections of the *Känguru-Wettbewerb* (2 in the medium and 2 in the difficult section). Panel B reports on results of a Mann-Whitney test. Outcome variable: number of omitted questions. We can reject the null hypothesis that the samples are drawn from the same distribution between all tested pairs in High School. p-values are reported in parentheses.

Gender gap in answering multiple-choice questions

Table 4.7: Raw Treatment Effects

	Overall (Q1-Q14)		Easy (Q1-Q5)		Medium (Q6-Q10)		Difficult (Q11-Q14)	
Panel A: Regression	*Vocational School*	*High School*	*Vocational School*	*Vocational School*	*Vocational School*	*High School*	*Vocational School*	*High School*
Treatments								
Incentivized	0.088	-0.406*	0.014	0.145**	0.074	-0.199*	-0.001	-0.062
	[0.120]	[0.247]	[0.045]	[0.076]	[0.046]	[0.115]	[0.046]	[0.095]
Female	0.165*	0.990***	0.041	0.078	0.094**	0.376**	0.029	0.536***
	[0.097]	[0.295]	[0.047]	[0.072]	[0.078]	[0.187]	[0.032]	[0.102]
Female × Incentive	-0.263**	-0.716**	-0.067	-0.007	-0.140***	-0.270	-0.056	-0.440***
	[0.116]	[0.364]	[0.057]	[0.094]	[0.054]	[0.210]	[0.042]	[0.136]
Controls								
SchoolFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Covariates	No	No	No	No	No	No	No	No
N	1220	870	1220	870	1220	870	1220	870
Panel B: Contrasts	*Gender Gap*							
Not Incentivized	0.165*	0.990***	0.041	0.077	0.094**	0.376**	0.029	0.536***
	[0.097]	[0.295]	[0.047]	[0.072]	[0.048]	[0.187]	[0.032]	[0.102]
Incentivized	-0.099	0.274	-0.026	0.071	-0.045*	0.106	-0.027	0.097
	[0.065]	[0.206]	[0.033]	[0.066]	[0.025]	[0.083]	[0.027]	[0.095]

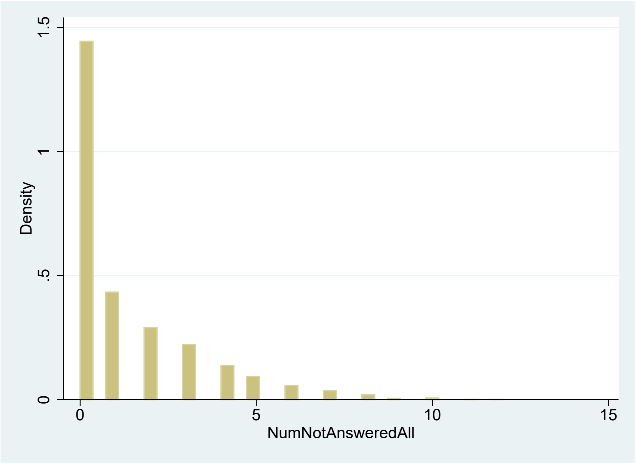
Note: Panel A reports the results of least squares regressions without covariates separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. The gender gap in skipping questions is captured by *Female* (Female=0: boys; Female=1: girls). Panel B reports the gender difference in skipping questions for incentivized and non-incentivized pupils resulting from Panel A. Dependent variable: number of skipped questions. Robust standard errors are reported in parentheses and clustered on classroom-level. * p<0.10, ** p<0.05, *** p<0.01

Table 4.8: Average Number of Omitted Questions, Probability of Success and Points achieved by Gender, Difficult Levels and by Incentivized

		Omitted questions (in percent)				Probability of success (in percent)				Aver. Points per Question			
		Easy	*Medium*	*Difficult*	*Overall*	*Easy*	*Medium*	*Difficult*	*Overall*	*Easy*	*Medium*	*Difficult*	*Overall*
Not Incentivized													
Vocational School	Male (N = 202)	3.37	8.66	7.67	5.50	39.70	24.50	46.53	38.32	0.55	0.24	1.72	0.74
	Female (N = 160)	4.38	13.44	9.38	7.50	36.69	21.88	45.94	35.88	0.44	0.06	1.59	0.61
High School	Male (N = 137)	12.41	20.15	20.80	17.57	63.60	46.13	41.12	52.18	1.34	1.05	1.19	1.17
	Female (N = 99)	13.94	27.07	34.09	24.39	60.54	45.15	36.53	49.55	1.22	0.90	0.58	0.92
Incentivized													
Vocational School	Male (N = 474)	4.05	11.92	7.59	6.56	38.84	30.49	52.43	40.13	0.53	0.46	1.99	0.84
	Female (N = 384)	3.44	9.38	5.99	5.32	37.47	24.22	50.26	37.78	0.48	0.22	1.88	0.73
High School	Male (N = 368)	9.29	15.43	18.95	14.25	62.79	38.78	35.76	47.12	1.36	0.78	0.85	1.01
	Female (N = 266)	11.13	17.97	22.09	16.70	57.46	36.53	36.06	44.56	1.14	0.69	0.84	0.89

Note: This table reports the percentage number of skipped question (*Omitted questions*), the probability to answer a question correct—the share of correct answers on all given answers—(*Probability of success*) and the average number of points per question (*Aver. Points per Question*) for each section differentiated by school type, gender and incentive groups.

Figure 4.3: Distribution of Omitted Questions



Field Data

Table 4.9: Correct Answers by Academic Year

Dep. Var:	*Overall*			*Easy Section*			*Medium Section*			*Difficult Section*		
Correct Answers	Boys	Girls	Difference	Boys	Girls	Difference	Boys	Girls	Difference	Boys	Girls	Difference
Grade												
3	40.23	38.57	-1.66***	63.75	60.59	-3.16***	41.63	40.16	-1.46***	15.30	14.95	-0.35
	(57440)	(49088)		(57440)	(49088)		(57440)	(49088)		(57440)	(49088)	
4	50.60	49.25	-1.34***	74.34	71.20	-3.14***	56.99	56.43	-0.56*	20.46	20.14	-0.33
	(62324)	(56373)		(62324)	(56373)		(62324)	(56373)		(62324)	(56373)	
5	40.64	38.02	-2.62***	64.58	62.23	-2.35***	39.71	36.48	-3.24***	17.64	15.36	-2.28***
	(77888)	(76884)		(77888)	(76884)		(77888)	(76884)		(77888)	(76884)	
6	47.29	45.20	-2.09***	70.60	69.39	-1.21***	47.88	45.35	-2.53***	23.39	20.86	-2.53***
	(78514)	(75714)		(78514)	(75714)		(78514)	(75714)		(78514)	(75714)	
7	32.38	30.84	-1.54***	45.20	41.45	-3.75***	33.23	31.47	-1.76***	18.70	19.60	0.90***
	(49118)	(44960)		(49118)	(44960)		(49118)	(44960)		(49118)	(44960)	
8	38.21	35.89	-2.32***	53.36	48.88	-4.48***	39.59	36.53	-3.06***	21.69	22.27	0.58*
	(35858)	(30912)		(35858)	(30912)		(35858)	(30912)		(35858)	(30912)	
9	41.57	39.06	-2.50***	68.71	66.44	-2.27***	35.74	33.39	-2.35***	20.25	17.36	-2.89***
	(25649)	(19773)		(25649)	(19773)		(25649)	(19773)		(25649)	(19773)	
10	45.89	41.96	-3.93***	73.46	69.64	-3.82***	41.27	37.08	-4.19***	22.94	19.16	-3.78***
	(15110)	(10601)		(15110)	(10601)		(15110)	(10601)		(15110)	(10601)	
11	43.89	38.69	-5.19***	71.62	67.26	-4.36***	40.94	33.45	-7.49***	19.10	15.37	-3.73***
	(6671)	(4006)		(6671)	(4006)		(6671)	(4006)		(6671)	(4006)	
12	44.21	38.12	-6.09***	72.05	66.51	-5.54***	40.84	32.65	-8.19***	19.73	15.19	-4.54***
	(1884)	(1060)		(1884)	(1060)		(1884)	(1060)		(1884)	(1060)	

Note: This table reports the means of correct answers in the *Känguru-Wettbewerb* test 2013. Cell entries report percentages and the number of pupils that participated in the test is reported in parentheses. *Difference* is the difference between girls - boys. The significance of gender differences is estimated by testing on the equality of proportions.

Table 4.10: Omitted and Correct Answers by Prize Winner

	Non Prize Winner						*Prize Winner*					
	Boys	Girls	Difference	Boys	Girls	Difference	Boys	Girls	Difference	Boys	Girls	Difference
Grade	*Omitted Answers*			*Correct Answers*			*Omitted Answers*			*Correct Answers*		
3	16.21 (53731)	17.09 (46639)	0.88^{***}	38.19 (53731)	37.00 (46639)	-1.19^{***}	9.34 (3709)	11.09 (2449)	1.75^{**}	69.60 (3709)	68.46 (2449)	-1.14
4	11.08 (58471)	11.58 (53661)	0.50^{***}	48.66 (58471)	47.75 (53661)	-0.91^{***}	4.51 (3853)	5.59 (2712)	1.08^{**}	79.90 (3853)	78.99 (2712)	-0.90
5	15.43 (72286)	15.99 (73618)	0.56^{***}	38.23 (72286)	36.58 (73618)	-1.65^{***}	6.89 (5602)	8.07 (3266)	1.18^{**}	71.80 (5602)	70.51 (3266)	-1.29
6	11.60 (73261)	11.50 (72401)	-0.10	44.92 (73261)	43.62 (72401)	-1.30^{***}	3.37 (5253)	4.46 (3313)	1.09^{**}	80.57 (5253)	79.69 (3313)	-0.88
7	20.95 (46002)	21.87 (43101)	0.92^{***}	30.32 (46002)	29.53 (43101)	-0.79^{***}	14.10 (3116)	15.74 (1859)	1.64	62.70 (3116)	61.11 (1859)	-1.59
8	17.84 (33467)	19.32 (29713)	1.48^{***}	35.89 (33467)	34.55 (29713)	-1.34^{***}	9.47 (2391)	10.47 (1199)	1.00	70.72 (2391)	69.27 (1199)	-1.45
9	22.01 (23873)	23.75 (19139)	1.74^{***}	39.37 (23873)	38.04 (19139)	-1.33^{***}	12.70 (1776)	13.61 (634)	0.91	71.31 (1776)	70.10 (634)	-1.21
10	21.85 (1411)	23.51 (10272)	1.66^{***}	43.72 (1411)	40.89 (10272)	-2.83^{***}	9.27 (999)	10.44 (329)	1.17	76.46 (999)	75.47 (329)	-0.99
11	24.20 (6187)	26.72 (3895)	2.52^{***}	41.63 (6187)	37.77 (3895)	-3.86^{***}	9.81 (484)	10.78 (111)	0.97	72.68 (484)	71.02 (111)	-1.65
12	25.99 (1741)	29.60 (1039)	3.61^{**}	41.67 (1741)	37.40 (1039)	-4.27^{**}	9.42 (143)	6.03 (21)	-3.39	75.06 (143)	73.65 (21)	-1.41

Note: This table reports the means of correct and omitted answers in the *Känguru-Wettbewerb* test 2013 by prize winners. Cell entries report percentages and the number of pupils that participated in the test is reported in parentheses. *Difference* is the difference between girls - boys. The significance of gender differences is estimated by testing on the equality of proportions.

Robustness Check

Table 4.11: Robustness Check - All Questions

	Negative Binomial		OLS	
	Vocational School	*High School*	*Vocational School*	*High School*
Treatments				
Incentivized	0.116	-0.388	0.110	-0.488**
	[0.084]	[0.250]	[0.111]	[0.244]
Female	0.112	0.982***	0.101	0.922***
	[0.104]	[0.408]	[0.0969]	[0.281]
Female × Incentive	-0.210	-0.747**	-0.218*	-0.684*
	[0.125]	[0.439]	[0.120]	[0.353]
Controls				
SchoolFE	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes
N	1193	867	1193	867

Note: This tables reports the results of a negative binomial regression and least squares regression over all questions (Q1-Q14) separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. The gender gap in skipping questions is captured by *Female* (Female=0: boys; Female=1: girls). Dependent variable: number of skipped questions. Covariates: last midterm grade, number of books at home, math curiosity (self-reported on 1-5 scale), academic year (grade 5 or 6). Robust standard errors are reported in parentheses and clustered on classroom-level. * p<0.10, ** p<0.05, *** p<0.01

Table 4.12: Robustness Check - Easy Questions

	Negative Binomial		OLS	
	Vocational School	*High School*	*Vocational School*	*High School*
Treatments				
Incentivized	0.033	-0.140*	0.032	-0.141*
	[0.039]	[0.080]	[0.043]	[0.079]
Female	0.025	0.070	0.021	0.074
	[0.047]	[0.112]	[0.049]	[0.074]
Female × Incentive	-0.060	-0.007	-0.058	-0.011
	[0.056]	[0.123]	[0.058]	[0.095]
Controls				
SchoolFE	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes
N	1193	867	1193	867

Note: This tables reports the results of a negative binomial regression and least squares regression over easy questions (Q1-Q5) separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. The gender gap in skipping questions is captured by *Female* (Female=0: boys; Female=1: girls). Dependent variable: number of skipped questions. Covariates: last midterm grade, number of books at home, math curiosity (self-reported on 1-5 scale), academic year (grade 5 or 6). Robust standard errors are reported in parentheses and clustered on classroom-level. * p<0.10, ** p<0.05, *** p<0.01

Table 4.13: Robustness Check - Middle Questions

	Negative Binomial		OLS	
	Vocational School	*High School*	*Vocational School*	*High School*
Treatments				
Incentivized	0.071	-0.259**	0.071	-0.264**
	[0.038]	[0.114]	[0.048]	[0.110]
Female	0.069	0.273**	0.075	0.295*
	[0.048]	[0.169]	[0.049]	[0.176]
Female × Incentive	-0.116**	-0.204	-0.122**	-0.226
	[0.059]	[0.182]	[0.058]	[0.202]
Controls				
SchoolFE	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes
N	1193	867	1193	867

Note: This tables reports the results of a negative binomial regression and least squares regression over medium questions (Q6-Q10) separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. The gender gap in skipping questions is captured by *Female* (Female=0: boys; Female=1: girls). Dependent variable: number of skipped questions. Covariates: last midterm grade, number of books at home, math curiosity (self-reported on 1-5 scale), academic year (grade 5 or 6). Robust standard errors are reported in parentheses and clustered on classroom-level. * p<0.10, ** p<0.05, *** p<0.01

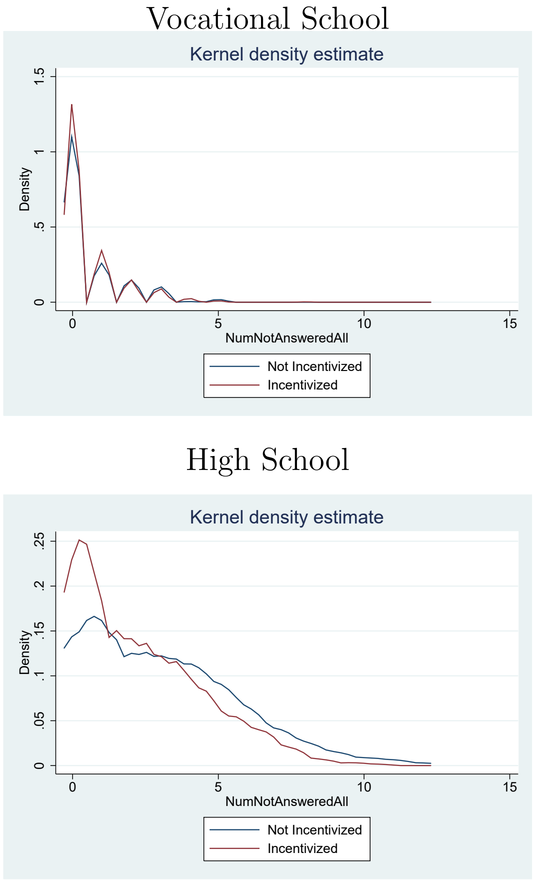
Table 4.14: Robustness Check - Difficult Questions

	Negative Binomial		OLS	
	Vocational School	*High School*	*Vocational School*	*High School*
Treatments				
Incentivized	0.009	-0.066	0.007	-0.083
	[0.035]	[0.100]	[0.040]	[0.095]
Female	0.004	0.573***	0.004	0.553***
	[0.040]	[0.164]	[0.030]	[0.099]
Female × Incentive	-0.038	-0.468***	-0.038	-0.446***
	[0.048]	[0.178]	[0.041]	[0.135]
Controls				
SchoolFE	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes
N	1193	867	1193	867

Note: This tables reports the results of a negative binomial regression and least squares regression over difficult questions (Q11-Q14) separately for High Schools and Vocational Schools linked by seemingly unrelated estimations and including school fixed effects. The gender gap in skipping questions is captured by *Female* (Female=0: boys; Female=1: girls). Dependent variable: number of skipped questions. Covariates: last midterm grade, number of books at home, math curiosity (self-reported on 1-5 scale), academic year (grade 5 or 6). Robust standard errors are reported in parentheses and clustered on classroom-level. * p<0.10, ** p<0.05, *** p<0.01

Kernel density estimation

Figure 4.4: Kernel Density Estimation Incentive vs. No Incentive

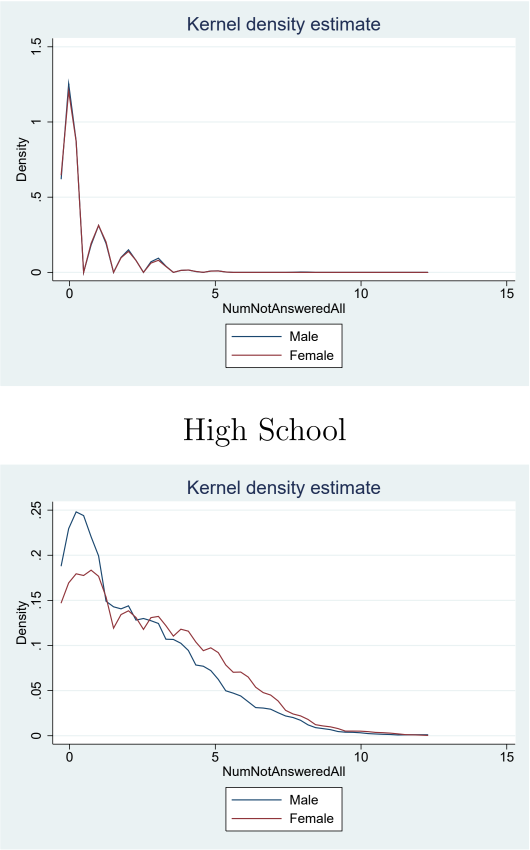


Vocational School

High School

Note: Figure (a) presents kernel density estimates for the number of skipped test questions for incentivized and non-incentivized pupils in Vocational Schools. Figure (b) presents kernel density estimates for the number of skipped test questions for incentivized and non-incentivized pupils in High Schools.

Figure 4.5: Kernel Density Estimation Male vs. Female



Vocational School

High School

Note: Figure (a) presents kernel density estimates for the number of skipped test questions for boys and girls in Vocational Schools. Figure (b) presents kernel density estimates for the number of skipped test questions for boys and girls in High Schools.

Chapter 5

Concluding Remarks

This thesis applies insights from behavioral economics to the educational sector and analyzes pupils' performance in multiple-choice tests by using the methodology of field experiments. Chapter 2 and Chapter 3 investigate how pupils can be incentivized to exert more (cognitive) effort in a mathematical test. Chapter 4 analyzes structural biases in multiple-choice testing formats.

The second chapter considers the signaling value of non-monetary rewards in secondary schools and analyzes to whom pupils want to reveal their educational achievements as well as the impact of rewards on test performance. We find that pupils with lower math grades prefer more often to signal their academic achievement to their parents (choice of the parents letter) and that pupils with higher math grades tend to signal their achievements to their peers (choice of the medal). On test performance, we find differences on the working of rewards for pupils who differ in their socio-economic background. While pupils in Vocational School seem not to be affected by rewards (treatment estimates are positive but not significant), pupils in High School decrease significantly their performance if rewards are predetermined. However, when allowing for choice over the incentive, we do not observe a decrease in pupils' performance. Hence, when designing non-monetary incentives in schools, educators should opt for a choice over incentives rather than predetermine rewards. Moreover, pupils' socio-economic background should be taken into account. Pupils from higher socio-economic strata need higher powered incentives than families of lower socio-economic strata.

The third chapter focuses on the effectiveness of gain and loss framing on the test performance of elementary pupils. If find that loss framing—pupils start with the maximum score—and a downward shift of the point scale—pupils start with a negative score—increase the number of correct answers compared to pupils who are graded "traditionally". This increase seems to be driven by two different mechanisms. While pupils in the loss framing condition increase the number of correct answers because they take more risky decisions, pupils in the negative endowment condition increase the number of correct answers because they answer more accurately. Moreover, the two treatment conditions work differently for high- and low-achieving pupils. While the former increase their performance under both conditions, loss framing is detrimental for the test performance of low-achieving pupils. However, there is no detrimental effect on low-achievers in the negative endowment condition. Considering the increasing number of behavioral insight teams in international institutions and governments, my results recommend not to implement loss framing in schools although it might be appealing to policy-makers due to its easy to implement and cost-effective characteristics. Instead, policy-makers should rather consider manipulations of the point scale, e.g. a downward shift.

To summarize the findings of Chapter 2 and Chapter 3, motivation and hence performance in mathematical tests can be changed by providing non-monetary incentives and by framing manipulations. More importantly, these changes are due to an effort effect and not due to more learning because the preparatory material did not prepare pupils in terms of the test content and because interventions were scheduled immediately before pupils had to take the test. Hence, interventions aiming at increasing pupils' attitude towards school, i.e. motivation, could be as important as interventions aiming at increasing pupils' learning behavior.

The fourth chapter analyzes whether multiple-choice testing formats favor answering strategies of certain groups in the population. We find evidence that pupils in High School and Vocational School differ in their answering strategies. Pupils in High School tend to skip more questions than pupils in Vocational School but pupils in High School score higher in the test than pupils in Vocational School. Moreover, we find a gender bias in answering multiple-choice questions among High School pupils. Girls tend to skip more answers than boys but only if questions are difficult, indicating that girls may suffer from a stereotype-threat. Furthermore, skipping more test items decreases test performance for pupils in High Schools. Complementing our experimental data with data of a nationwide test shows that these gender differences seem to exist across all grades (grade 3 - grade 12).

To conclude, this thesis shows that motivation is a key input to succeed in the educational system and reports on important results for policy-makers aiming at implementing behavioral concepts like loss framing and non-monetary incentives in schools. However, the experiments conducted in this thesis have some limitations, treatment effects can only be interpreted for the populations studied and we can infer only on short-run effects. It needs further research to study pupils' behavior in repeated interventions. Moreover, it would be of interest to conduct the experiments in a high-stakes testing environment, although low-stakes testing environments give already valuable insights because they allow to exclude an overlapping incentive effect stemming from a high-stakes testing environment. Furthermore, the negative impact of incentives on pupils in High School found in Chapter 2 and the negative results of loss framing on low-performers in Chapter 3 highlight the value of randomized field experiments and show that more field experiments should be conducted *before* changing school laws or the institutional setting as it might be costly to *not* experiment (List 2011).

The methodology of field experiments in the economics of education offers a range of opportunities for future research, inter alia in combination with the increased use of new digital learning technologies. Randomized field experiments combined with e-learning devices and teaching software are potentially cost-effective, subjects are unaware of participating in an experiment and, more importantly, they could help to better track and understand how learning is processed. Furthermore, it is also important to learn more about how institutional decision makers can be convinced to participate in large scale field experiments. In particular, it would be interesting to know which features of the intervention have to be made salient; is it the research question, the "investments" to be made by the institution (i.e. time, personnel capacities) or can extrinsic (financial and non-financial) incentives increase the likelihood of participation? Moreover, it would be valuable to know if characteristics of the schools' community such as the distance to the next university, the unemployment rate or the political attitude increase the likelihood of a school to participate (this is ongoing research with Gerhard Riener, Heinrich-Heine-Universität Düsseldorf and Sebastian Schneider, Georg-August-Universität Göttingen).

> *"For instance, when a nemesis claims that this experiment will cost the firm too much money, I often respond that we are "costing" the firm too much money by not experimenting."*
>
> *— John List, 2011, Journal of Economic Perspectives, pp. 12*

Bibliography

George Akerlof and Rachel Kranton. Identity and the Economics of Organizations. *Journal of Economic Perspectives*, 19:9–32, 2005.

Saziye Akyol, James Key, and Kala Krishna. Hit or Miss? Test Taking Behavior in Multiple Choice Exams. Working Paper 22401, National Bureau of Economic Research, July 2016.

Ingvild Almås, Alexander Cappelen, Kjell Salvanes, Erik Sorensen, and Bertil Tungodden. What Explains the Gender Gap in College Track Dropout? Experimental and Administrative Evidence. *The American Economic Review*, 106(5):296–302, 2016.

Carole Ames. Classrooms: Goals, Structures, and Student Motivation. *Journal of Educational Psychology*, 84(3):261–271, 1992.

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.

Simon Calmar Andersen, Louise Voldby Beuchert-Pedersen, Helena Skyt Nielsen, and Mette Kjærgaard Thomsen. The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial. Working paper, 2015.

Ola Andersson, Håkan Holm, Jean-Robert Tyran, and Erik Wengström. Risk Aversion Relates to Cognitive Ability: Preferences or Noise? *Journal of the European Economic Association*, 14(5):1129–1154, 2016.

Silke Anger and Daniel Schnitzlein. Cognitive Skills, Non-Cognitive Skills, and Family Background: Evidence from Sibling Correlations. Discussion Paper 9918, Institute for the Study of Labor (IZA), April 2016.

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

Joshua Angrist and Alan Krueger. Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Joshua Angrist and Victor Lavy. The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. *The American Economic Review*, 99:1384–1414, 2009.

Heather Antecol, Ozkan Eren, and Serkan Ozbeklik. The Effect of Teacher Gender on Student Achievement in Primary School. *Journal of Labor Economics*, 33(1): 63–89, 2015.

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.

Dan Ariely, Anat Bracha, and Stephan Meier. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *The American Economic Review*, 99(1):544–55, March 2009.

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.

Orley Ashenfelter, Colm Harmon, and Hessel Oosterbeek. A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias. *Labour Economics*, 6(4):453–470, 1999.

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

David Austen-Smith and Roland Fryer. An Economic Analysis of “Acting White”. *The Quarterly Journal of Economics*, 120:551–583, 2005.

David Autor, Lawrence Katz, and Melissa Kearney. Trends in U.S. Wage Inequality: Revising the Revisionists. *Review of Economics and Statistics*, 90(2):300–323, May 2008.

David Autor, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman. Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes. Working Paper 22267, National Bureau of Economic Research, May 2016.

Francesco Avvisati, Marc Gurgand, Nina Guyon, and Eric Maurin. Getting Parents Involved: A Field Experiment in Deprived Schools. *The Review of Economic Studies*, 81(1):57–83, 2014.

Ghazala Azmat, Caterina Calsamiglia, and Nagore Iriberri. Gender Differences in Response to Big Stakes. *Journal of the European Economic Association*, forthcoming, 2016.

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

Brad Barber and Terrance Odean. Boys will be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics*, 116(1): 261–292, 2001.

Jürgen Baumert and Anke Demmrich. Test Motivation in the Assessment of Student Skills: The Effects of Incentives on Motivation and Performance. *European Journal of Psychology of Education*, 16(3):441–462, 2001.

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.

Gershon Ben-Shakhar and Yakov Sinai. Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies. *Journal of Educational Measurement*, 28(1):23–35, 1991.

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

Daniel Benjamin, Sebastian Brown, and Jesse Shapiro. Who is "Behavioral"? Cognitive Ability and Anomalous Preferences. *Journal of the European Economic Association*, 11(6):1231–1255, 2013.

Eric Bettinger. Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics*, 94(3):686–698, 2012.

Eric Bettinger and Robert Slonim. Patience Among Children. *Journal of Public Economics*, 91(1–2):343–363, February 2007.

Julian Betts. Chapter 7 - The Economics of Tracking in Education. In Stephen Machin Eric Hanushek and Ludger Wößmann, editors, *Handbook of the Economics of Education*, volume 3, pages 341–381. Elsevier, 2011.

Manudeep Bhuller, Magne Mogstad, and Kjell Salvanes. Life Cycle Earnings, Education Premiums and Internal Rates of Return. Working Paper 20250, National Bureau of Economic Research, June 2014.

Maria Bigoni, Margherita Fort, Mattia Nardotto, and Tommaso Reggiani. Cooperation or Competition? A Field Experiment on Non-monetary Learning Incentives. *The B.E. Journal of Economic Analysis and Policy*, 15(4):1753–1792, 2015.

Sandra Black, Paul Devereux, and Kjell Salvanes. Staying in the Classroom and out of the Maternity Ward? The Effect of Compulsory Schooling Laws on Teenage Births. *The Economic Journal*, 118(530):1025–1054, 2008.

Moussa Blimpo. Team Incentives for Education in Developing Countries: A Randomized Field Experiment in Benin. *American Economic Journal: Applied Economics*, 6(4):90–109, 2014.

Pedro Dal Bo, Andrew Foster, and Louis Putterman. Institutions and Behavior: Experimental Evidence on the Effects of Democracy. *The American Economic Review*, 100(5):2205–2229, December 2010.

Niall Bolger and Thomas Kellaghan. Method of Measurement and Gender Differences in Scholastic Achievement. *Journal of Educational Measurement*, 27(2): 165–174, 1990.

Alison Booth and Patrick Nolen. Gender Differences in Risk Behaviour: Does Nurture Matter? *The Economic Journal*, 122(558):56–78, 2012.

Lex Borghans, James Heckman, Bart Golsteyn, and Huub Meijers. Gender Differences in Risk Aversion and Ambiguity Aversion. *Journal of the European Economic Association*, 7(2-3):649–658, 2009.

Wilfried Bos, Heike Wendt, Olaf Köller, and Christoph Selter. *TIMSS 2011 Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Waxmann Verlag, New York, NY, 2012.

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.

Giorgio Brunello and Martin Schlotter. Non-Cognitive Skills and Personality Traits: Labour Market Relevance and Their Development in Education & Training Systems. Discussion Paper 5743, Institute for the Study of Labor (IZA), May 2011.

Stephen Burks, Jeffrey Carpenter, Lorenz Goette, and Aldo Rustichini. Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment. *Proceedings of the National Academy of Sciences*, 106(19):7745–7750, 2009.

Justine Burns, Simon Halliday, and Malcolm Keswell. Gender and Risk Taking in the Classroom. SALDRU Working Papers 87, Southern Africa Labour and Development Research Unit, University of Cape Town, 2012.

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

Bettina Büttner and Stephan Thomsen. Are We Spending too Many Years in School? Causal Evidence of the Impact of Shortening Secondary School Duration. *German Economic Review*, 16(1):65–86, 2015.

Rachel Caffyn. Attitudes of British Secondary School Teachers and Pupils to Rewards and Punishments. *Educational Research*, 31(3):210–220, 1989.

Colin Camerer and Robin Hogarth. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, 19(1-3):7–42, 1999.

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.

Scott Carrell, Richard Fullerton, and James West. Does Your Cohort Matter? Measuring Peer Effects in College Achievement. *Journal of Labor Economics*, 27(3): 439–464, 2009.

Scott Carrell, Marianne Page, and James West. Sex and Science: How Professor Gender Perpetuates the Gender Gap. *The Quarterly Journal of Economics*, 125 (3):1101–1144, 2010.

Shelly Chaiken and Yaacov Trope. *Dual-Process Theories in Social Psychology.* Guilford Press, 1999.

Gary Charness, Ramn Cobo-Reyes, Natalia Jimenez, Juan Lacomba, and Francisco Lagos. The Hidden Advantage of Delegation: Pareto Improvements in a Gift Exchange Game. *The American Economic Review*, 102(5):2358–379, August 2012.

John Chelonis, Rebecca Flake, Ronald Baldwin, Donna Blake, and Merle Paule. Developmental Aspects of Timing Behavior in Children. *Neurotoxicology and Teratology*, 26(3):461–476, 2004.

Arnaud Chevalier, Peter Dolton, and Melanie Lührmann. “Making It Count”: Evidence from a Field Study on Assessment Rules, Study Incentives and Student Performance. Discussion Paper 8582, Institute for the Study of Labor (IZA), October 2014.

Katherine Baldiga Coffman. Evidence on Self-Stereotyping and the Contribution of Ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660, 2014.

Kerris Cooper and Kitty Stewart. Does Money Affect Children’s Outcomes? Case reports, Centre for Analysis of Social Exclusion, LSE, 2013.

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. *The American Economic Review*, 97(2):31–47, 2007.

David Cutler and Adriana Lleras-Muney. Education and Health: Evaluating Theories and Evidence. Working Paper 12352, National Bureau of Economic Research, July 2006.

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.

Thomas Deckers, Armin Falk, Fabian Kosse, and Hannah Schildberg-Hörisch. How Does Socio-Economic Status Shape a Child's Personality? Discussion Paper 8977, Institute for the Study of Labor (IZA), April 2015.

Hartmut Ditton. Schulübertritte, Geschlecht und soziale Herkunft. In Hartmut Ditton, editor, *Kompetenzaufbau und Laufbahnen im Schulsystem*, pages 63–87. Waxmann, Münster, 2007. ISBN 978-3-8309-1887-5.

Thomas Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Are Risk Aversion and Impatience Related to Cognitive Ability? *The American Economic Review*, 100(3):1238–1260, June 2010.

Paul Dolan and Matteo Galizzi. Getting Policy-Makers to Listen to Field Experiments. *Oxford Review of Economic Policy*, 30(4):725–752, 2014.

Martin Dowson and Dennis McInerney. Psychological Parameters of Students' Social and Work Avoidance Goals: A Qualitative Investigation. *Journal of Educational Psychology*, 93(1):35–42, 2001.

Angela Duckworth and Martin Seligman. Self-Discipline gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores. *Journal of Educational Psychology*, 98(1):198–208, 2006.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Chapter 61 - Using Randomization in Development Economics Research: A Toolkit. In Paul Schultz and John Strauss, editors, *Handbook of Development Economics*, volume 4, pages 3895–3962. North-Holland, Elsevier, 2007.

Christian Dustmann. Parental Background, Secondary School Track Choice, and Wages. *Oxford Economic Papers*, 56(2):209–230, 2004.

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

David Dwyer and Jeffrey Hecht. Minimal Parental Involvement. *School Community Journal*, 2(2):53–66, 1992.

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.

Seymour Epstein. Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, 49(8):709–724, 1994.

Mara Paz Espinosa and Javier Gardeazabal. Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment. *Journal of Economics and Management*, 9(2):107–135, 2013.

Armin Falk and James Heckman. Lab Experiments are a Major Source of Knowledge in the Social Sciences. *Science*, 326(5952):535–538, 2009.

Leon Feinstein, Ricardo Sabates, Tashweka Anderson, Annik Sorhaindo, and Cathie Hammond. Chapter 4 - What are the effects of education on health? In *Measuring the Effects of Education on Health and Civic Engagement: Proceedings of the Copenhagen Symposium*, pages 171–354. OECD, Paris, 2006.

Nicole Fortin, Philip Oreopoulos, and Shelley Phipps. Leaving Boys Behind. *Journal of Human Resources*, 50(3):549–579, 2015.

Shane Frederick. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

Norman Frederiksen. The Real Test Bias: Influences of Testing on Teaching and Learning. *American Psychologist*, 39(3):193–202, 1984.

Bruno Frey. How Intrinsic Motivation is Crowded Out and In. *Rationality and Society*, 6(3):334–352, 1994.

Bruno Frey and Reto Jegen. Motivation Crowding Theory. *Journal of Economic Surveys*, 15(5):589–611, 2001.

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

Roland Fryer. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, 31:373–427, 2013.

Roland Fryer and Steven Levitt. An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, 2(2):210–240(31), 2010.

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.

Roland Fryer, Steven Levitt, and John List. Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights. Working Paper 21477, National Bureau of Economic Research, August 2015.

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.

Naomi Gafni and Estela Melamed. Differential Tendencies to Guess as a Function of Gender and Lingual-Cultural Reference Group. *Studies in Educational Evaluation*, 20(3):309–319, 1994.

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

Uri Gneezy, Kenneth Leonard, and John List. Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society. *Econometrica*, 77(5):1637–1664, 2009.

Uri Gneezy, Stephan Meier, and Pedro Rey-Biel. When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives*, 25(4):191–209, 2011.

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

Claudia Goldin, Lawrence Katz, and Ilyana Kuziemko. The Homecoming of American College Women: The Reversal of the College Gender Gap. *Journal of Economic Perspectives*, 20(4):133–156, 2006.

Joshua Goodman. The Labor of Division: Returns to Compulsory Math Coursework. Working paper, HKS Faculty Research Working Paper Series RWP12-032, John F. Kennedy School of Government, Harvard University, 2012.

Cornelia Gresch, Jürgen Baumert, and Kai Maaz. Empfehlungsstatus, Übergangsempfehlung und der Wechsel in die Sekundarstufe I: Bildungsentscheidungen und soziale Ungleichheit. In Jürgen Baumert, Kai Maaz, and Ulrich Trautwein, editors, *Bildungsentscheidungen*, pages 230–256. VS Verlag für Sozialwissenschaften, 2010.

Wayne Grove and Tim Wasserman. Incentives and Student Learning: A Natural Experiment with Economics Problem Sets. *The American Economic Review*, 96 (2):447–452, May 2006.

Eric Hanushek. The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24:1141–1177, 1986.

Eric Hanushek and Ludger Wößmann. Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation. *Journal of Economic Growth*, 17(4):267–321, 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.

Colm Harmon and Ian Walker. Estimates of the Economic Return to Schooling for the United Kingdom. *The American Economic Review*, 85(5):1278–1286, 1995.

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

Glenn Harrison and John List. Field Experiments. *Journal of Economic Literature*, 42(4):1009–1055, December 2004.

James Heckman, Rodrigo Pinto, and Peter Savelyev. Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *The American Economic Review*, 103(6):2052–2086, October 2013.

Sara Heller, Anuj Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold Pollack. Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. *The Quarterly Journal of Economics*, forthcoming, 2016.

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.

Caroline Hoxby. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics*, 115(4):1239–1285, 2000.

Janet Hyde and Janet Mertz. Gender, Culture, and Mathematics Performance. *Proceedings of the National Academy of Sciences*, 106(22):8801–8807, 2009.

Janet Hyde, Sara Lindberg, Marcia Linn, Amy Ellis, and Caroline Williams. Gender Similarities Characterize Math Performance. *Science*, 321(5888):494–495, 2008.

Alex Imas, Sally Sadoff, and Anya Samek. Do People Anticipate Loss Aversion? *Management Science*, accepted, 2016.

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.

Peter Jensen and Astrid Würtz Rasmussen. The Effect of Immigrant Concentration in Schools on Native and Immigrant Children's Reading and Math Skills. *Economics of Education Review*, 30(6):1503–1515, 2011.

Robert Jensen. The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, 125(2):515–548, 2010.

Kathrin Jonkmann, Kai Maaz, Marko Neumann, and Cornelia Gresch. Übergangsquoten und Zusammenhänge zu familiärem Hintergrund und schulischen Leistungen. Deskriptive Befunde. 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 123–150. Bundesministerium für Bildung und Forschung, Referat Bildungsforschung, Bonn, 2010.

Štěpán Jurajda and Daniel Münich. Gender Gap in Performance under Competitive Pressure: Admissions to Czech Universities. *The American Economic Review*, 101 (3):514–518, 2011.

Daniel Kahneman. Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, 93(5):1449–1475, 2003.

Daniel Kahneman and Shane Frederick. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In Thomas Gilovich, Dale Griffin, and Daniel Kahneman, editors, *Heuristics and biases: The psychology of intuitive judgment*, volume 49-81. New York: Cambridge University Press, 2002.

Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–292, 1979.

Meredith Kimball. A New Perspective on Women's Math Achievement. *Psychological Bulletin*, 105(2):198–214, 1989.

Nand Kishor and Maureen Godfrey. The Effect of Information Framing on Academic Task Completion. *Educational Psychology*, 19(1):91–101, 1999.

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

Michael Kosfeld and Susanne Neckermann. Getting More Work for Nothing? Symbolic Awards and Worker Performance. *American Economic Journal: Microeconomics*, 3(3):86–99, August 2011.

Michael Kremer, Esther Duflo, and Pascaline Dupas. Peer Effects, Teacher Incentives, and the Impact of Tracking. *The American Economic Review*, 101: 1739–1774, 2011.

Sebastian Kube, Michel André Maréchal, and Clemens Puppea. The Currency of Reciprocity: Gift Exchange in the Workplace. *The American Economic Review*, 102(4):1644–1662, 2012.

Nicola Lacetera and Mario Macis. Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2):225–237, 2010.

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.

Victor Lavy. Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110(6):1286–1317, 2002.

Thomas Lemieux. Postsecondary Education and Increasing Wage Inequality. *The American Economic Review*, 96(2):195–199, 2006.

Edwin Leuven, Hessel Oosterbeek, and Bas Klaauw. The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment. *Journal of the European Economic Association*, 8(6):1243–1265, 2010.

Steven Levitt and John List. Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review*, 53(1):1–18, 2009.

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.

Elly-Ann Lindström and Erica Lindahl. The Effect of Mixed-Age Classes in Sweden. *Scandinavian Journal of Educational Research*, 55(2):121–144, 2011.

John List. Field Experiments: A Bridge Between Lab and Naturally Occurring Data. *The B.E. Journal of Economic Analysis and Policy*, 5(2):Article 8, 2007.

John List. Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off. *Journal of Economic Perspectives*, 25(3):3–15, 2011.

John List and Anya Savikhin Samek. The Behavioralist as Nutritionist: Leveraging Behavioral Economics to Improve Child Food Choice and Consumption. *Journal of Health Economics*, 39:135–146, 2015.

John List, Sally Sadoff, and Mathis Wagner. So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design. *Experimental Economics*, 14(4):439–457, 2011.

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

Lance Lochner and Enrico Moretti. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *The American Economic Review*, 94(1):155–189, 2004.

James Loewen, Phyllis Rosser, and John Katzman. Gender Bias in SAT Items. *Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA*, 1988.

George Loewenstein. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin*, 116(1):75–98, 1994.

Stephen Machin, Olivier Marie, and Sunčica Vujić. The Crime Reducing Effect of Education. *The Economic Journal*, 121(552):463–484, 2011.

Anandi Mani, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. Poverty Impedes Cognitive Function. *Science*, 341(6149):976–980, 2013.

Milton Marquis, Bharat Trehan, and Wuttipan Tantivong. The Wage Premium Puzzle and the Quality of Human Capital. *International Review of Economics & Finance*, 33:100–110, 2014.

Paul McCoubrie. Improving the Fairness of Multiple-Choice Questions: A Literature Review. *Medical Teacher*, 26(8):709–712, 2004.

Patrick McEwan. Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, 85(3):353–394, 2015.

Philip Mellizo, Jeffrey Carpenter, and Peter Hans Matthews. Workplace Democracy in the Lab. *Industrial Relations Journal*, 45(4):313–328, 2014.

Kevin Milligan, Enrico Moretti, and Philip Oreopoulos. Does Education Improve Citizenship? Evidence from the United States and the United Kingdom. *Journal of Public Economics*, 88(9-10):1667 – 1695, 2004.

Karin Monstad, Carol Propper, and Kjell Salvanes. Education and Fertility: Evidence from a Natural Experiment. *Scandinavian Journal of Economics*, 110(4): 827–852, 2008.

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

Karthik Muralidharan and Venkatesh Sundararaman. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1):39–77, 2011.

Derek Neal. Chapter 6 - The Design of Performance Pay in Education. In Stephen Machin Eric Hanushek and Ludger Wößmann, editors, *Handbook of The Economics of Education*, volume 4 of *Handbook of the Economics of Education*, pages 495–550. Elsevier, 2011.

Susanne Neckermann and Bruno Frey. And the Winner Is!? The Motivating Power of Employee Awards. *The Journal of Socio-Economics*, 46:66–77, 2013.

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

Muriel Niederle. Chapter - 8, Gender. In John Kagel and Alvin Roth, editors, *Handbook of Experimental Economics*, volume 2, pages 481–562. Princeton: Princeton University Press, 2016.

Muriel Niederle and Lise Vesterlund. Do Women Shy Away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics*, 122(3):1067–1101, 08 2007.

Muriel Niederle and Lise Vesterlund. Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, 24(2):129–144, 2010.

Muriel Niederle and Alexandra Yestrumskas. Gender Differences in Seeking Challenges: The Role of Institutions. Working Paper 13922, National Bureau of Economic Research, April 2008.

Asako Ohinata and Jan Van Ours. How Immigrant Children Affect the Academic Achievement of Native Dutch Children. *The Economic Journal*, 123(570):308–331, 2013.

Philip Oreopoulos. Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter. *The American Economic Review*, 96(1):152–175, 2006.

Philip Oreopoulos. Do Dropouts Drop out Too Soon? Wealth, Health and Happiness from Compulsory Schooling. *Journal of Public Economics*, 91(11):2213–2229, 2007.

Philip Oreopoulos and Kjell Salvanes. Priceless: The Nonpecuniary Benefits of Schooling. *Journal of Economic Perspectives*, 25(1):159–184, 2011.

Evren Ors, Frédéric Palomino, and Eloïc Peyrache. Performance Gender Gap: Does Competition Matter? *Journal of Labor Economics*, 31(3):443–499, 2013.

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.

Wiebke Paulus and Hans-Peter Blossfeld. Schichtspezifische Präferenzen oder sozioökonomisches Entscheidungskalkül? Zur Rolle elterlicher Bildungsaspirationen im Entscheidungsprozess beim Übergang von der Grundschule in die Sekundarstufe. *Zeitschrift für Pädagogik*, 53(4):491–508, 2007.

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.

Jan Retelsdorf and Jens Möller. Entwicklungen von Lesekompetenz und Lesemotivation Schereneffekte in der Sekundarstufe? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 40(4):179–188, 2008.

Steven Rivkin, Eric Hanushek, and John Kain. Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458, 2005.

Richard Ryan and Edward Deci. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology*, 25(1):54–67, 2000.

Bruce Sacerdote. Peer Effects in Education: How might they Work, How big are they and How much do we Know Thus Far? In Erik Hanushek, Stephen Machin, and Ludger Wößmann, editors, *Handbook of the Economics of Education*, volume 3, pages 249–277. Elsevier, June 2011.

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

Perihan Ozge Saygin. Do Girls Really Outperform Boys in Educational Outcomes? Working Paper 14-05, University of Mannheim, Department of Economics, 2014.

Martin Schlotter, Guido Schwerdt, and Ludger Wößmann. Econometric Methods for Causal Evaluation of Education Policies and Practices: A Non-Technical Guide. *Education Economics*, 19(2):109–137, 2011.

Andrei Shleifer. Psychologists at the Gate: A Review of Daniel Kahneman's "Thinking, Fast and Slow". *Journal of Economic Literature*, 50(4):1080–1091, 2012.

Selcuk Sirin. Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, 75(3):417–453, 2005.

Matthew Springer, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, Lockwood, Daniel McCaffrey, Matthew Pepper, and Brian Stecher. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. *Society for Research on Educational Effectiveness*, 2011.

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.

Heinrich Stumpf and Julian Stanley. Gender-Related Differences on the College Board's Advanced Placement and Achievement Tests, 1982–1992. *Journal of Educational Psychology*, 88(2):353–364, 1996.

Matthias Sutter, Stefan Haigner, and Martin Kocher. Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations. *The Review of Economic Studies*, 77(4):1540–1566, 2010.

Matthias Sutter, Martin Kocher, Daniela Glätzle-Rüetzler, and Stefan Trautmann. Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior. *The American Economic Review*, 103(1):510–531, 2013.

Pinchas Tamir. Positive and Negative Multiple Choice Items: How Different are they? *Studies in Educational Evaluation*, 19(3):311–325, 1993.

Daniel Tannenbaum. Do Gender Differences in Risk Aversion Explain the Gender Gap in SAT Scores? Uncovering Risk Attitudes and the Test Score Gap.[PRELIMINARY]. 2012.

The European Commission. *Tackling the gender pay gap in the European Union*. Publications Office of the European Union, Luxembourg, 2014. ISBN 978-92-79-36068-8.

The US Bureau of Labour Statistics. Women in the Labor Force: A Databook. Report 1052, 2014.

Philip Trostel, Ian Walker, and Paul Woolley. Estimates of the Economic Return to Schooling for 28 Countries. *Labour Economics*, 9(1):1–16, 2002.

Jean Twenge and Keith Campbell. Self-Esteem and Socioeconomic Status: A Meta-Analytic Review. *Personality and Social Psychology Review*, 6(1):59–71, 2002.

Timothy Urdan and Martin Maehr. Beyond a Two–Goal Theory of Motivation and Achievement: A Case for Social Goals. *Review of Educational Research*, 65(3): 213–243, 1995.

Simon Veenman. Cognitive and Noncognitive Effects of Multigrade and Multi-Age Classes: A Best-Evidence Synthesis. *Review of Educational Research*, 65(4):319–381, 1995.

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

Li-fang Zhang and Gerard Postiglione. Thinking Styles, Self-Esteem, and Socio-Economic Status. *Personality and Individual Differences*, 31(8):1333–1346, 2001.

Daniel John Zizzo. Experimenter Demand Effects in Economic Experiments. *Experimental Economics*, 13(1):75–98, 2010.

Eidesstattliche Versicherung

Ich, Valentin Wagner, versichere an Eides statt, dass die vorliegende Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der “Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf” erstellt worden ist.

Düsseldorf, der 14. Oktober 2016

Unterschrift