# Essays in Behavioral Economics of Education: Experimental and Empirical Studies on Information, Beliefs, and Educational Decisions

Inauguraldissertation

zur

Erlangung des Doktorgrades

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Universität zu Köln

2018

vorgelegt von

Mira Fischer, M.A.

aus

Aachen

Referent: Prof. Dr. Dirk Sliwka

Korreferent: Prof. Dr. Bernd Irlenbusch

Tag der Promotion: 14.03.2018

# Acknowledgments

First and foremost I would like to thank my PhD advisor, Dirk Sliwka, for supporting me and for allowing me to grow as a researcher during these past five years. The hard work and long discussions on our project have allowed me to learn how to conduct experimental research and to appreciate theory. He has given me the freedom to follow my own interests in various projects and it has been a greatly enriching experience to work under his guidance. I am also grateful for him being a role model as a sincere and committed researcher who is curious, and deeply interested in the questions he studies.

I would also like to thank the other members of my thesis committee, Bernd Irlenbusch, for support and advice over several years, and Matthias Heinz for being my primary source for practical advice on how to survive in academia.

I am grateful to Patrick Kampkötter and Valentin Wagner, co-authors on projects contained in this thesis. At different stages of my PhD, they have taught me much hands-on knowledge related to empirical research and running field experiments and it was very motivating to have such dedicated co-authors.

I would like to thank Alex Bryson, John List, Alexander Cappelen and Bertil Tungodden for being great hosts during my research stays in London, Chicago and Bergen. All of these three stays have allowed me to experience exciting and stimulating new research environments and to gain many new insights and ideas.

I would also like to thank current and past colleagues: Gari Walkowitz for many open discussions about research and lots of other things. Lea Cassar for conversations on meditation and for trying to maneuver bureaucracy with me in order to run a field experiment. Anja Bodenschatz, Gönül Doğan, Florian Engl, Rainer Michael Rilke, Marina Schröder, Caroline Stein and Timo Vogelsang for thought-provoking discussions, advice, and support. Tobias Danzeisen, Lucas Grunwitz, Mirjam Reetz, Theresa Schwan, and Carolin Wegner for their great help with programming and running lab experiments.

I particularly want to thank my partner Rogier for his love and support, for proof-reading this thesis, for endless conversations and for always being by my side. I would like to thank my grandfather Paul for always being interested in my intellectual development, be it in Philosophy or in Economics. Lastly, I would like to thank my parents for their unconditional love and for always enabling me to freely choose my path in life.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

This thesis investigates people's educational decisions. It consists of four research papers that apply a broad range of research methods to different educational settings. All papers have in common that they put students as decision-makers center stage and focus on how they incorporate information into their beliefs and behavior.

The classical economic approach to education assumes that people choose their investments in human capital to maximize their lifetime utility (Hanushek, 1979; Todd and Wolpin, 2003). Empirical evidence suggests, however, that people often make educational decisions that may not benefit them in the long run. For example, many students drop out of education but later regret having done so (Bridgeland et al., 2006) or procrastinate on preparation even for very important exams (Steel, 2007). Such "mistakes" in one's educational decisions may imply large individual and social costs as educational attainment is a strong predictor for happiness (Oreopoulos and Salvanes, 2011), health (Silles, 2009; Buckles et al., 2016), and behavior towards others (Milligan et al., 2004; Heckman et al., 2006).

Levitt et al. (2016) have pointed out that it is important to do "basic research" in economics of education that is able to identify single factors in the educational process in order to inform policy making and educational interventions, and behavioral economics may have particularly much to contribute to this endeavor (Lavecchia et al., 2016). While some of the best known economic field experiments targeting educational outcomes have studied the effectiveness of (monetary and non-monetary) incentives (Angrist and Lavy, 2009; Kremer et al., 2009; Fryer, 2011; Bettinger,

2012; Levitt et al., 2016), with mixed results, education economists in recent years have increasingly focused on factors influencing the process of human capital formation that are not captured by the classical economic approach, such as beliefs, preferences and character traits (Heckman and Kautz, 2012). For example, whether someone is motivated to do their best at school, to pursue a college degree or to apply for a demanding job may depend on their beliefs about their abilities (Benabou and Tirole, 2002, 2016; Heckman et al., 2006), their perceived benefits of education (Reuben et al., 2017), their risk aversion (Davies et al., 2002), their conscientiousness and openness to new experience (Noftle and Robins, 2007), and their preference for competition (Niederle and Vesterlund, 2010).While individual characteristics, such as gender and family background are known to be correlated with, for example, people's beliefs about their abilities (Filippin and Paccagnella, 2012) and their labor market expectations (Dawson, 2017; Reuben et al., 2017), the mechanisms underlying these differences are not well understood. The aim of the studies comprised in this thesis is to contribute to their understanding and the promising field of behavioral economics of education.

Economists have recently started to focus on the formation of students' beliefs (Alan et al., 2016; Kosse et al., 2016) as well as on the effects of social comparisons (Tran and Zeckhauser, 2012; Azmat et al., 2016) in education. Following these two strands of research, the articles in this thesis focus on how beliefs about academic ability or the quality of education and relative performance evaluations affect students' educational decisions. The scope of this thesis reaches from an empirical investigation, providing quasi-experimental evidence on the effects of the German universities Excellence Initiative on ability sorting and perceived educational quality, and a randomized field experiment in secondary schools, testing the effects of different types of relative performance information on high-stakes educational outcomes, to economic laboratory experiments, investigating the psychological mechanisms underlying the motivation to invest in human capital. The different chapters are summarized below.

Chapter 2 investigates potential spillover effects of the German Excellence Initiative on university education. It is co-authored with Patrick Kampkötter and is forthcoming as "Effects of German Universities' Excellence Initiative on Ability Sorting of Students and Perceptions of Educational Quality" in the Journal of Institutional and Theoretical Economics. Using data from a nationally representative student survey commissioned by Germany's Federal Ministry of Education and Research, we apply a difference-in-differences strategy to study the effects of this excellence competition on the quality of a university's enrollments and it's perceived quality of education. In the first part of the paper, we study the effects of increased differentiation in research reputation and research funding on ability sorting of students among universities. The announcements of the winning institutions of the Excellence Initiative are rare and highly publicized events in which information on the universities that are considered the best research universities in the country suddenly becomes common knowledge. Thus, they are suitable for studying whether a university's reputation has an effect on its success in recruiting talented students. We find that the award of excellence status allows a university to enroll significantly better students in three subsequent admissions terms, which increases differences in student ability between "excellent" and "non-excellent" universities.

In the second part of the paper, we study an important factor of enrollment decisions – the perceived quality of a university's education – by analyzing whether a signal of research quality influences students' perceptions of educational quality, as measured by their satisfaction ratings. We are able to study how students' perceptions respond to the award of the label itself because students were surveyed immediately after universities received excellence status and before research money tied to it could be used for organizational changes. Our results show a positive and highly significant effect of the excellence label on the students' perceptions of quality of education and, consistently, on perceived job opportunities after graduation. We also find that none of the items referring to the students' satisfaction with their personal life that are unrelated to their university show any significant response to the award of the label.

This indicates that improvements in a university's student ratings due to the label occur not because students identify with an "excellent" institution (and the positive emotions this might involve) but because students update their beliefs about the quality of their university's vis-à-vis other (non-excellent) universities' education. However, when students are surveyed three years later, student ratings largely return to previous levels, although the universities still enjoy excellence status. Overall, we find that the research competition resulted not only in stronger competition for (and more inequality of) research funds, which was its declared aim, but also in a more unequal distribution of talented students across universities, an effect that has been found to contribute to increasing wage inequality among graduates (see, e.g., Hoxby and Terry, 1999; Bergh and Fink, 2009). The excellence status seems to attract more students to apply to an institution because it is perceived as a signal of high educational quality and, consequently, better job prospects. As universities have limited capacity and high school grades generally are the most important selection criterion, "excellent" universities can have more competitive admissions. Our results thus shed light on an important side-effect of competition policies for public universities.

Chapter 3 studies the effects of a randomized feedback intervention on high-stakes educational outcomes in a field experiment and is co-authored with Valentin Wagner. In order to study the role of reference frame and timing for the effectiveness of feedback, students aged 11-12 years in 19 classes in 7 secondary schools received private written feedback from their teachers. The feedback notes either contained (i) information about their absolute rank in the last math exam, (ii) information about their change in ranks between the two previous math exams, or (iii) no information. Students received the feedback either (a) 1-3 days or (b) immediately before the last exam of the semester. Students were provided with relative performance information as people are strongly motivated by it, even in the absence of any tangible benefits (Charness and Rabin, 2002; Azmat and Iriberri, 2010). However, rank feedback that compares one's level of performance to one's peers' levels does

not capture individual progress well, especially when there are large ability differences within the same class. Feedback that compares students not in terms of their performance levels but in terms of their changes in performance might help to mitigate this problem while maintaining the motivational effects of social comparison. Additionally, feedback was given at different points in time because outcomes in the workplace or educational settings may be influenced by different types of effort exerted at different times. While earlier feedback may have a stronger impact on preparation efforts, feedback given more immediately before a task may potentially have a stronger effect on effort at the task itself. The timing of feedback may also matter if feedback influences both expectations and emotions and the latter have stronger effects on motivation in the short run than in the long run (Lempert and Phelps, 2014).

We find that feedback is only effective to increase subsequent performance when given a few days before the last exam, possibly by countering students' tendency to procrastinate and start preparations too late, and that both change and level feedback work equally well to increase performance. These effects are driven by boys and by students who recently suffered a decrease in their performance while differences in self-reported competitiveness do not explain behavior. In contrast, any feedback given to students immediately before the exam tends to lower subsequent performance but the overall effects are not significant. Our results give interesting insights into how relative performance feedback works in educational settings and has implications for the design of feedback in other situations where the ability to motivate people is crucial, such as the workplace or health care. Our findings indicate that relative feedback may be particularly motivating when one has recently got worse and should be given early enough such that one still has a chance to make up for it.

Chapter 4 investigates the effects of peer group ability on confidence in own ability in a laboratory experiment and is single-authored. Understanding how within-group and between-group information affect ability beliefs is crucial for settings in

which ability groups are deliberately formed to increase individual performance, such as the classroom and the workplace. Although there are literatures studying the effects of either within-group (Murphy and Weinhardt, 2014; Elsner and Isphording, 2017) or between-group (Coffman, 2014; Dee, 2014) information, it is not yet well explored how both interact to influence ability beliefs and motivation. The net effect of assignment to a weaker group (versus a stronger group) on confidence in own ability may be negative or positive, depending on how each type of information is interpreted. In our experimental setting, group assignment depends imperfectly on ability such that the ability distributions of the two groups overlap and the ability signal from group assignment is noisy. This generates randomness of group assignment that allows for the causal identification of the effect of group assignment on ability beliefs and on subsequent performance. We randomly vary whether subjects only receive information about their performance relative to their group or whether they learn additionally whether they were assigned to a weaker or a stronger group and that group assignment depends imperfectly on ability. This allows us to study the causal effects of assignment to a weaker or a stronger group, and its interaction with salience of the group assignment mechanism, on confidence in ability and on subsequent test outcomes.

Our results show that when the group assignment mechanism is non-salient, it does not matter for subjects' confidence whether they are assigned to the weaker or the stronger group, however, when the group assignment mechanism is salient, weaker group assignment makes people less confident. We also find that subjects are on average less confident when the group assignment mechanism is salient than when it is non-salient. This is found to be the case due to weaker group assignment making people more underconfident than stronger group assignment making people overconfident, indicating that people overweigh negative information as compared to positive information. When grouping is non-salient, subjects on average give quite correct estimates of their ability rank. However, when grouping is salient, subjects' beliefs are significantly decalibrated, indicating that people overweigh ability sig-

nals coming from between-group information. With respect to test outcomes, we find that salient ability grouping has a positive effect on the performance of lower ability individuals while it has a negative effect on the performance of higher ability individuals. This is driven by opposite effects for these groups when they are assigned to the weaker group. While the performance of lower ability individuals increases when learning they were assigned to the weaker group, the performance of higher ability individuals decreases when learning they were assigned to the weaker group. Overall, our results suggest that ability grouping may have negative effects on people's confidence in their ability and that the positive effect of worse peers on confidence if relative ability between groups is non-salient may be greatly outweighed by the negative effect of having worse peers when relative ability between groups is salient. In settings where ability grouping is done visibly, our results also suggest that forming ability groups may harm those people who are negatively surprised by weaker group assignment more than it may benefit those who are positively surprised by stronger group assignment. These findings may help to understand the effects of ability grouping in the field and may inform the design of educational and workplace settings.

Finally, Chapter 5 studies the causal effects of beliefs on motivation in a laboratory experiment and is co-authored with Dirk Sliwka. The key purpose of this paper is to distinguish two dimensions of confidence – confidence in one's level of prior knowledge and confidence in one's learning ability – and to study causal effects of changes in these dimensions of a person's confidence on investments in human capital. Reinforcement of confidence in these two dimensions likely has very different effects, as the first dimension is related to one's ex-ante probability of passing a test while the second one is related to how much one's passing probability increases when exerting learning efforts. We first illustrate these belief dimensions in a simple formal model and then study the effects of variations in both dimensions experimentally. To investigate the causal effects of the two dimensions of confidence, we exogenously vary feedback scores subjects receive about their performance in two

prior tests. One of these tests measures their prior knowledge, the other test measures their ability to memorize information. The random component in the feedback scores generates exogenous variation in the agents' confidence in the two dimensions, which we use as instrumental variables to estimate causal effects of confidence on investment decisions and test outcomes.

We find that a higher confidence in learning ability raises learning investments irrespective of the level of prior knowledge. Confidence in knowledge, however, has a negative effect on investments of individuals with above average prior knowledge and a positive effect on investments of individuals with below average prior knowledge. With respect to test outcomes, we find that raising the confidence in learning of individuals with below average prior knowledge improves their rank in the final test and their probability of passing it, however, we do not find a beneficial effect for individuals who already had above average prior knowledge. Mirroring the effects of confidence in knowledge on learning investments, we find that raising confidence in knowledge of individuals with above average prior knowledge decreases their outcomes in the final test whereas it has the opposite effect on individuals with below average prior knowledge. The motivational role of confidence has attracted substantial interest from different fields in economics in recent years. Our results may help to explain why confidence in one's abilities may sometimes be positively (Benabou and Tirole, 2002, 2003; Heckman et al., 2006) and sometimes negatively (Malmendier and Tate, 2005; Niederle and Vesterlund, 2007) related to a person's outcomes. Insights about the different effects of confidence in learning ability and confidence in prior knowledge have implications not only for the design of interventions aimed at positively affecting academic motivation but also for subjective performance evaluation policies in firms and other organizations. Our results imply that rater leniency when assessing someone's ability to acquire a certain skill or achieve a future outcome can be beneficial, while rater leniency with respect to past achievements can be detrimental.

Overall, the results from the four studies presented in this thesis contribute to

our understanding of how people incorporate different types of information into their beliefs and educational decisions and we may in particular gain the following insights, ranging from more practical to more theoretical: (1) Students respond to signals about a university's research quality when deciding at which institution to pursue a degree and for this reason policies intended to foster research competition between universities may have side-effects on universities' quality of admissions. A university's excellence status also influences student satisfaction ratings but the effect is transient. (2) High stakes educational outcomes may be influenced by giving feedback about past performance, especially if it recently decreased, however the timing of feedback is crucial and it should be given early enough. (3) It is important to distinguish between beliefs in different ability dimensions as fostering some ability beliefs may raise the motivation to learn while fostering others may lower it. (4) People may overweigh ability signals from between-group comparisons, especially when they are negative, and for this reason ability grouping may lead to a decalibration of ability beliefs.

The studies outlined above will be presented in detail in the following.

# Chapter 2

# Effects of German Universities' Excellence Initiative on Ability Sorting of Students and Perceptions of Educational Quality

Co-authored with Patrick Kampkötter[1]

## 2.1 Introduction

In the past two decades, intensified competition among universities for funds and students has been widely observable in many countries (The Economist, 2015) . In Europe, this competition is fostered by the Bologna process that began in 1999 and aims to render educational institutions and degrees more comparable and compatible. In its wake, many countries adopted policies to raise the quality of higher education and research by promoting a more efficient use of resources in public universities. Stronger competition for students has also resulted from the availability and increased prominence of a number of national and international university rankings in recent years. In 2005, in order to foster competition in research, the German federal government and federal states jointly launched the Excellence Initiative, a contest that promises substantial amounts of additional funds and the prestigious title of "university of excellence" to successful institutions. The aim of this contest is to strengthen academic research and international visibility by promoting compe-

[1]My co-author and I contributed equally to the design of the study, to the data analysis, and to writing the paper.

tition in research among universities. It consists of three lines of funding: graduate schools, "clusters of excellence" to promote interdisciplinary research on socially relevant topics, and, so-called "future concepts" (or "institutional strategies") - the most important line of funding - which are "aimed at developing top-level university research in Germany and increasing its competitiveness at an international level" (German Research Foundation, 2016a). To be eligible to compete for the "future concepts" line of funding, a university must have been granted funding for at least one graduate school and at least one cluster of excellence. The program had an initial budget of 1.9 billion euros for the three funding lines and an additional budget of 2.7 billion euros was granted for the second phase of the program starting in 2012 (German Research Foundation, 2016b). All funds are to be spent on research only. Universities who were successful in the "future concepts" line of funding were awarded the label "university of excellence" and subsequently received up to an additional 70 million euros over a five year period. In this paper, we focus on the "future concepts" line of funding as it was tied to the largest amounts of money and the label "university of excellence" and was only awarded to a small number of institutions. This label evidently brought these institutions considerable public attention , and they have used the label for public relations. Our aim is to test for two particular spillover effects from this competition on higher education. In the first part of the paper, we study the effects of increased differentiation in research reputation and research funding on ability sorting of students among universities. The announcements of the winning institutions of the Excellence Initiative are rare and highly publicized events in which information on the universities that are considered the best research universities in the country suddenly becomes common knowledge. Thus, they are suitable for studying whether a university's reputation has an effect on its success in recruiting talented students. We find that the award of excellence status allows a university to enroll significantly better students in three subsequent admissions terms, which increases differences in student ability between "excellent" and "non-excellent" universities. In the second part of the paper, we study an impor-

tant factor of enrollment decisions - the perceived quality of a university's education - by analyzing whether a signal of research quality influences students' perceptions of educational quality, as measured by their satisfaction ratings. We are able to study how students' perceptions respond to the award of the label itself because students were surveyed immediately after universities received excellence status and before research money tied to it could be used for organizational changes. Our results show a positive and highly significant effect of the excellence label on the students' perceptions of quality of education and, consequently, on perceived job opportunities after graduation. We also find that none of the items referring to the students' satisfaction with their personal life that are unrelated to their university show any significant response to the award of the label. This indicates that improvements in a university's student ratings due to the label occur not because students identify with an "excellent" institution (and the positive emotions this might involve) but because students update their beliefs about the quality of their university's vis-à-vis other (non-excellent) universities' education. However, when students are surveyed three years later, student ratings largely return to previous levels, although the universities still enjoy excellence status.

## 2.2 Related Literature

### 2.2.1 Determinants of Quality of Admissions

An important line of research in the economics of higher education focuses on the institutional factors influencing student choice. In particular, students are interested in how much they will enjoy attending a university and how much their education will earn them in the labor market. Hence, both expectations of personal experience and development (DesJardins and Toutkoushian, 2005) and of job opportunities (Schaafsma, 1976; Lazear, 1977) are important drivers of enrollment in higher education. Thus, higher education can be described as having both an experience and a credence good property. The experience good property derives from the fact that

students generally only know what it "feels" like to pursue a certain academic program at a certain university once they have already (at least partially) completed it. The credence good property derives from the non-transparency of educational production and students' uncertainty about the labor market's valuation of the human capital they acquire at a certain university. Generally, credence and experience good properties create a situation of asymmetric information, in which the producer knows more about the properties of a good than the consumer (Akerlof, 1970; Wolinsky, 1995; DesJardins and Toutkoushian, 2005). This situation creates a demand for expert advice - for example expressed by quality labels - that allows consumers to reduce their uncertainty about the properties of such a good (Dulleck and Kerschbamer, 2006). The decision to attend a particular university affects the course of a person's life and often poses a once-in-a-lifetime choice. These kinds of decisions are particularly difficult to make, which is why people tend to be bad at making them (Benartzi and Thaler, 2007). Hence, quality signals, such as a high rank or the award of a label, which are easier for better universities to acquire, may be used by prospective students as a signal of a university's quality and may guide their enrollment decisions. Indeed, there is robust evidence that the reputation of an institution reflected by its rank in a league table is an important factor in student choice (Hossler et al., 1989; Weiler, 1996; Abbott and Leslie, 2004; Mueller and Rockerbie, 2005) and particularly affects the matriculation probability of high-ability students (Griffith and Rask, 2007; Gibbons et al., 2015). Hoxby (2009) has shown that due to increased student mobility and decreased information costs, U.S. students' college preferences have become more responsive to resources and peers, resulting in stronger ability sorting between colleges. In the U.K., Broecke (2015)has found that a worsening of a university's rank leads to a small, but statistically significant reduction in the number of applications and in the quality of accepted applicants. In Germany, the factors affecting student choice have received little attention Obermeit (2012). Recent studies have focused on few subjects, such as medicine and pharmaceutics, for which there is centralized matching of students with institutions by the

clearing house for university admissions, and the role of distance between students' hometown and the nearest university in application decisions (Braun et al., 2010; Spiess and Wrohlich, 2010; Hüber and Kübler, 2011). Horstschräer (2012) has investigated how the application likelihood of high-ability students to medical schools is influenced by the Excellence Initiative and has found that becoming a "university of excellence" significantly increases the application likelihood of high-ability students. The first part of our analysis draws a more comprehensive picture of the effects of the Excellence Initiative than Horstschräer (2012) by covering students of all subjects of study and investigating changes in the actual composition of students over time. Additionally, Bruckmeier et al. (2014) is closely related to our study, and these authors show that the loss of excellence university status within the Excellence Initiative negatively affects the number of enrolled first-year students in the subsequent winter term at universities in the federal state of Baden-Wuerttemberg. They also present evidence that this result is driven by the loss in reputation due to the withdrawal of excellence status and not due to a decrease in university quality. Conversely, being awarded excellence status had no significant effect on enrollment quantity. Whereas Bruckmeier et al. (2014) focus on the number of newly enrolled graduates, we analyze the effects of the excellence initiative on ability sorting.

### 2.2.2 Determinants of Perceived Quality of Education

Since education is a credence and experience good, potential students are likely to use quality labels or rankings provided by external bodies to reduce information asymmetries. If research quality and educational quality are positively correlated, and evidence suggests that this is indeed the case (Ford et al., 1999; Dahl and Smimou, 2011), it is rational to interpret "excellence status" - although awarded to universities solely based on research merits - as a signal of educational quality. As students were surveyed in the same semester in which some universities received excellence status, and as the disbursement of research funds began later during that semester, any potential effects of the new status on student ratings are likely driven

by the label "university of excellence" and not by any institutional changes. Because students likely care little about research quality and a lot about educational quality when making their enrollment decision, students' belief in this correlation is assumed when analyzing the effect of the label on enrollment. Many studies in the field of consumer psychology have shown that labels affect beliefs about a product's non-observable properties (Teisl et al., 2008). However, we are not aware of any previous studies analyzing how a new signal about a university's research quality affects students' perceptions of educational quality. Showing that current students' ratings of educational quality respond to a label awarded for research will also help us to shed light on the psychological mechanism by which the research competition might affect the enrollment decisions of new students. A rationale for the existence of such an effect is that as students rate their university on a given scale, they implicitly rate it relative to other universities with which they have little or no experience. When their institution receives a label they interpret as revealing information about the institution's high educational quality relative to other institutions, they update their belief about the relative quality of the institution's education and rate it higher on the given scale, although no actual changes have taken place. One can distinguish between experience-related factors (ratings of teaching, course content, supervision, acquired skills, etc.) and expectations-related factors (expected labor market outcomes) of perceived quality of education. There is evidence from the U.S. that job opportunities are significantly better and starting salaries are significantly higher for graduates of more respected institutions (Black et al., 2005). We thus also expect students' labor market expectations to respond to the label: first, because higher perceived quality of education implies better perceived acquired qualifications, and second, because students may hold the belief that the label also independently affects potential employers' expectations with respect to the quality of graduates. Our analysis of responses in students' perceptions will be organized according to this distinction between experience-related factors and expectations-related factors and will focus on common items typically used in student surveys.

## 2.3 Data

We use data from a national student survey administered by the University of Konstanz on behalf of Germany's Federal Ministry of Education and Research. The data set comprises a representative sample of German students in tertiary education and covers 18 universities and 15 polytechnics (Fachhochschulen). Twelve waves of data were collected between winter semester 1982/1983 and winter semester 2012/2013, although not all 33 institutions are included in all waves as some institutions were included later and data collection in other institutions was discontinued. The data are collected every two to three years from a new random sample of students at covered institutions, with approximately 8,000 students per wave (Simeaner et al., 2013). The data set is representative of students at German universities and polytechnics with respect to attributes such as gender, subject of study, and age and institutions were selected to guarantee a representative coverage of federal states (Multrus, 2004). In winter semester 2012/2013, the last available wave, the response rate amounted to 18.6 percent. The survey data consist of information on student characteristics, including university attended, field of study, type of degree program, number of semesters, admission to a program during a summer or a winter term, full-time or a part-time student status, and demographics such as gender, age, and parents' highest level of education. The data also contain information on the grade point average (GPA) of the Abitur, the German high school diploma, which is a measure of a student's academic ability that is still the most important admission criterion for the vast majority of programs at German universities. Furthermore, information is available on a large number of items measuring student attitudes and satisfaction, such as ratings of content, supervision, acquired skills, and practical relevance of education, as well as expected labor market outcomes (see Table 2.11 in the Appendix for a description of the survey items). In our analysis, we use data on full-time and part-time students who enrolled after 1990 contained in 7 waves collected in winter semesters 1994/1995 through 2012/2013, i.e., the waves surrounding the first, second and third round of the Excellence Initiative. There are two sur-

vey waves coinciding with the first (2006) and third (2012) round of the Excellence Initiative. We restrict the data set to universities and exclude polytechnics because only the former were eligible to participate in the excellence competition. We also restrict the sample to universities that are present in at least three different waves. This leaves us with a total of approximately 37,000 students enrolled at 15 different universities . The data set contains information on two successful universities from the first round, one successful university from the second round, and one successful university from the third round of the competition: "university of excellence" status was announced for the University of Karlsruhe and the University of Munich (LMU) on October 13, 2006, for the University of Freiburg on October 19, 2007, and for the Technical University of Dresden on June 15, 2012. Descriptive statistics are shown in Table 2.5 in the Appendix. The average proportion of female students in our sample is 55 percent, the average number of semesters is 6.6, and the average high school diploma GPA is 2.2. The majority of the students are enrolled in the humanities and the social sciences.

## 2.4 Quality of Admissions

### 2.4.1 Empirical Strategy

To examine whether becoming a "university of excellence" affects the competitiveness of admissions (and student demand for a given university) in subsequent admission terms, we use the average high school GPA of newly enrolled students as the dependent variable. We estimate the following baseline specification of an OLS regression model:

$$GPA(z-score)_{ijt} = \alpha + \beta Excellent(A)_{jt} + \gamma University_j + \delta Cohort_t + \zeta IndividualControls_{ijt} + \varepsilon_{ijt} \quad (2.1)$$

where $GPA(z-score)_{ijt}$ is the standardized school GPA of student i who enrolled at university j in year t. We standardize grades over the entire sample to zero mean and unit variance to abstract from the German grading scale (1.0 = excellent, 4.0 = sufficient, less than 4.0 = fail) and to make the effect sizes internationally comparable. The $Excellent(A)_{jt}$ dummy is equal to 1 for all the students who enrolled (= were in their first semester) in a university after the university was labeled excellent and is equal to 0 otherwise. We include fixed effects for university to control for time-constant heterogeneity among universities and fixed effects for cohort to control for time-varying heterogeneity constant over universities, both potentially influencing the competitiveness of admissions. Since $Excellent(A)_{jt}$ varies within the awarded universities (Dresden, Freiburg Karlsruhe, and Munich) over time cohorts and stays constant in the non-awarded universities, this dummy, given university and time fixed effects, identifies the difference-in-differences effect of the award of excellence status on admissions. Furthermore, we include the following individual-level control variables: age, gender, parents' level of education, field of study, full-time or part-time student status, degree program (e.g., bachelor's, master's, state examination, Diplom), and whether the student was admitted during the summer term. The degree program dummies allow us to control for the gradual conversion from the former German system to the international system of bachelor's and master's programs during the Bologna process. The summer term admission dummy allows us to identify students who did not enroll during the main winter term admissions and instead enrolled during summer term admissions. Summer term admissions account for 14.3 percent of total admissions in our sample and might have different admission criteria. In a further specification, we interact $Excellent(A)_{jt}$ with separate dummies for the years following the competition to account for time trends in the selectivity of universities after receiving excellence status. For example, $Excellent(A)_{jt} * 1^{st}Year_{jt}$ identifies students who enrolled during the first year (summer or winter semester) after the university was awarded excellence status. This specification allows us to investigate when the effect begins and whether or after

how much time it wanes. To investigate whether the selectivity of universities was more responsive to excellence status in some fields than in others, we include interaction effects between $Excellent(A)_{jt}$ and $FieldOfStudy_{ijt}$ in a further specification. This allows us to investigate whether certain fields of study drive the response of admissions to the award of the label. For all the specifications, we present results both with and without controls for the presence of tuition fees ($TuitionFee(A)_{jt}$) and double high school graduation cohorts ($DoubleCohort(A)_{jt}$) in some German federal states at the time of admission. We consider it important to test whether our results are robust to these reforms because both of them might have affected the number of applicants at universities and hence, the competitiveness of admissions. The presence of tuition fees at some universities might drive students to apply to universities in other federal states without tuition fees or might affect the transition from high school to university (Dwenger et al., 2012; Hübner, 2012; Bruckmeier et al., 2013; Bruckmeier and Wigger, 2014). The presence of a double cohort in a federal state likely drives up the number of applicants at universities located in that federal state. For all the regressions, standard errors clustered on university level are reported.

### 2.4.2 Results

Table 2.1 and Table 2.2 contain OLS regression results estimating the impact of the Excellence Initiative on the quality of admissions, measured by the GPA of the students' high school diploma. In our baseline regression, five cohorts after the first wave of the Excellence Initiative are included. Column (1) of Table 2.1 presents the results from our baseline regression with standardized GPA. The coefficient of the excellence dummy (Excellent(A)) is negative and statistically significant. Note that in the German grading system, a smaller grade is a better grade. The results indicate that in the six years following the award of university of excellence status, a university's admissions were, on average, 0.125 standard deviations better than the admissions of universities without the excellence label. This is a sizeable effect

compared to the between university difference in grades and comparable in size to the effects of randomized controlled interventions in higher education. Long-term field experiments in schools report similar effect sizes (Angrist et al., 2006; Fryer, 2014). These results are also consistent with evidence showing a sorting of more able students into higher quality education institutions (Black et al., 2005). In column (2), we also control for tuition fees and double cohorts, which only slightly decreases the coefficient of interest. Column (3) presents the regression results for the interaction between the excellence dummy and six dummies identifying each year since the receipt of the award, again with additional controls for tuition fees and double cohorts in column (4). The results reveal that the overall effect of the award of excellence status on admissions is driven by the first three years after the award. This can also be observed by looking at the change in raw average grades before and after the award (see Figure A1 in the Appendix). The positive effect on admissions seems slightly larger in the second and third years than in the first year: however, Wald tests show that only the coefficients of the first and the third year are significantly different from each other ($\beta 1^{st}Year = \beta 2^{nd}Year: \ p = 0.198$; $\beta 1^{st}Year = \beta 3^{rd}Year: \ p = 0.053$; $\beta 2^{nd}Year = \beta 3^{rd}Year: \ p = 0.633$). After the third year, the effect seems to wane. The negative (but insignificant) interaction coefficient identifying the 6th year after the original award is a weak indication that the renewal of excellence status, similar to the original award, has a positive (but noisier) effect on admissions. In principle, the effect of excellence status on admissions could be driven by universities' restricting their capacities in the years after the award in order to become more "elite" and allow only a handful of students with very good GPAs to enroll. However, legal regulations prevent public universities in Germany from freely adjusting their capacity. Rather, the education ministries of the federal states determine how many places for new enrollment each university has to supply each semester. This means that a change in the competitiveness of admissions is driven by student demand for places at a given university. The evidence thus suggests that the effects of the Excellence Initiative on overall admissions are

driven by an increase in medium-term student demand for places and that it is the novelty of the excellence status (and the media attention it entails) rather than the status alone that allows universities to recruit better students.

Table 2.1: Excellence Status and Quality of Admissions

| Dependent variable: GPA Abitur (standardized) | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Excellent | –0.125** | –0.100** | | |
| | (0.0489) | (0.0453) | | |
| Excellent × 1st year | | | –0.0961** | –0.0813** |
| | | | (0.0340) | (0.0366) |
| Excellent × 2nd year | | | –0.183** | –0.157* |
| | | | (0.0758) | (0.0735) |
| Excellent × 3rd year | | | –0.204** | –0.175** |
| | | | (0.0695) | (0.0644) |
| Excellent × 4th year | | | 0.0140 | 0.0299 |
| | | | (0.0750) | (0.0764) |
| Excellent × 5th year | | | 0.0340 | 0.0696 |
| | | | (0.0927) | (0.0798) |
| Excellent × 6th year | | | –0.0863 | –0.0551 |
| | | | (0.0677) | (0.0738) |
| Tuition fees | | –0.0431** | | –0.0413** |
| | | (0.0168) | | (0.0161) |
| Double cohort | | –0.0073 | | –0.0360 |
| | | (0.0547) | | (0.0571) |
| Observations | 38,904 | 38,904 | 38,904 | 38,904 |
| Adjusted $R^2$ | 0.173 | 0.173 | 0.173 | 0.173 |

*Note:* We regress school GPA (standardized with zero mean and unit variance over the entire sample) on a dummy that indicates whether a student enrolled in a university after the university was labeled excellent. In columns (3) and (4), this dummy is separated into six dummies for each year following the award of excellence status. Columns (2) and (4) also control for tuition fees and double cohorts. All the regressions contain a constant and cohort and university fixed effects. Additionally, all the regressions control for field of study, degree program, summer-term admissions, part-time study, age, gender, and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To further investigate whether certain study subjects are driving the identified effect of the excellence status on admissions, we interact the excellence dummy with dummies for different fields of study. As observed in Table 2.2, enrollment in economics responds most strongly to the award of excellence status , with student ability significantly improving more than half a standard deviation, followed by en-

rollment in medicine, law, and the social sciences (compared to the baseline group humanities). A considerably weaker response to the excellence status is detectable for admissions in the natural sciences. It is, however, unlikely that the stronger competitiveness of admissions in economics is the reason why these admissions respond more strongly to the excellence label than admissions in the natural sciences because, as observed in the coefficients of the field of study dummies, economics students on average have a worse GPA than students in the natural sciences . Additionally, the effect cannot be explained by the label revealing more information about the quality of research in economics than in the natural sciences, as the excellence universities in our sample qualified to compete for the third line of funding (and the label) because they all had won excellence funds for graduate schools and research clusters (only) in the natural sciences (and none in economics). However, the difference in response to the label might be driven by economics students' placing more weight than students in the natural sciences on the alleged benefits of attending an excellent university, for example with respect to labor market signaling. Further analyses of the items asking about motivation to choose a certain program or university support this rationale: the economics students were more concerned about their earnings prospects when choosing a program and attached greater importance to a university's "tradition and reputation" when choosing at which university to study than the natural sciences students. To rule out that any one university alone is driving our results, we also run robustness checks excluding each excellence university in turn, which does not alter the results. We also rule out that summer term admissions are driving our results. (See Tables 2.6 and 2.7 in the Appendix.) Overall, our results suggest that there is a significant and sizeable medium-run effect of the Excellence Initiative on ability sorting at German universities, that this effect is strongest for economics students, and that "excellent" universities are able to recruit better school leavers at the expense of universities that did not succeed in this competition for three years after the award of excellence status. However, we do not find evidence that successful universities benefit in terms of better enrollments in the longer run.

Table 2.2: Quality of Admissions - Interaction with Field of Study

| Dependent variable: | | |
|---|---|---|
| GPA Abitur (standardized) | (1) | (2) |
| Excellent | 0.0829 | 0.110** |
| | (0.0502) | (0.0482) |
| Excellent × Social sciences | –0.240*** | –0.241*** |
| | (0.0393) | (0.0398) |
| Excellent × Law | –0.330*** | –0.329*** |
| | (0.0454) | (0.0446) |
| Excellent × Economics | –0.537*** | –0.539*** |
| | (0.113) | (0.113) |
| Excellent × Medicine | –0.325*** | –0.324*** |
| | (0.0319) | (0.0314) |
| Excellent × Natural sciences | –0.176** | –0.178** |
| | (0.0592) | (0.0605) |
| Excellent × Engineering | –0.147 | –0.153 |
| | (0.132) | (0.133) |
| Excellent × Other | –0.102 | –0.104 |
| | (0.171) | (0.171) |
| Social sciences | –0.0056 | –0.0052 |
| | (0.0499) | (0.0497) |
| Law | –0.125*** | –0.125*** |
| | (0.0349) | (0.0352) |
| Economics | 0.0670 | 0.0678 |
| | (0.0570) | (0.0570) |
| Medicine | –0.444*** | –0.443*** |
| | (0.0502) | (0.0503) |
| Natural sciences | –0.140*** | –0.140*** |
| | (0.0311) | (0.0310) |
| Engineering | 0.0832* | 0.0838* |
| | (0.0439) | (0.0439) |
| Other | 0.125* | 0.124* |
| | (0.0661) | (0.0661) |
| Tuition fees | | –0.0431** |
| | | (0.0170) |
| Double cohort | | –0.0034 |
| | | (0.0566) |
| Observations | 38,904 | 38,904 |
| Adjusted $R^2$ | 0.174 | 0.174 |

*Note:* We regress school GPA (standardized with zero mean and unit variance over the entire sample) on interaction terms between a dummy that indicates whether a student enrolled in a university after the university was labeled excellent and dummies for the field of study. Humanities is the reference category. Column (2) also controls for tuition fees and double cohorts. Both regressions contain a constant and cohort and university fixed effects. Additionally, all the regressions control for degree program, summer-term admissions, part-time study, age, gender, and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

## 2.5 Perceived Quality of Education

### 2.5.1 Empirical Strategy

To investigate whether the award of excellence status immediately affects students' perceptions of the quality of an institution's education, we study the relationship between recently having been named a "university of excellence" and an institution's student evaluations. We estimate the following baseline specification of an ordered logit model:

$$StudentEvaluation_{ijt} = \alpha + \beta Excellent(B)_{jt} + \gamma Excellent(B)_{jt+1} + \delta University_j + \zeta Wave_t + \eta IndividualControls_{ijt} + \varepsilon_{ijt} \quad (2.2)$$

where $StudentEvaluation_{ijt}$ denotes different survey items measuring student i's evaluation of the educational quality of university j, which she is attending at the time of survey wave t. The items are chosen to match criteria for student satisfaction used by internationally known university rankings such as the CHE ranking, the Times Higher Education World University Rankings, the Academic Ranking of World Universities (Shanghai Ranking), or the U.S. News & World Report's college rankings. (See Table 2.11 in the Appendix for a precise definition of each item.) To ensure comparability between the different item scales in the regression models, the items are standardized to zero mean and unit variance. The $Excellent(B)_{jt}$ dummy is equal to 1 if a rating was given by a student in the winter semester immediately after the university in which he or she is enrolled was awarded excellence status and is equal to 0 otherwise. The $Excellent(B)_{jt+1}$ dummy identifies the ratings of students at universities with excellence status collected in the following survey wave (3 years later). We include fixed effects for university and survey wave to control for time-constant heterogeneity among universities and time-varying heterogeneity constant over universities potentially influencing student ratings. Since $Excellent(B)_{jt}$

varies within the awarded universities over the survey waves and remains constant in the non-awarded universities, this dummy, given university and wave fixed effects, identifies the difference-in-differences effect of the award of the excellence label on student ratings before the research funds tied to the award could be used for organizational changes. Longer-term effects of the "excellence label" cannot be cleanly identified because the research funds tied to the label could have caused actual changes. We thus focus on the short-term effects of excellence status on student satisfaction ratings. However, it is still interesting to see whether student ratings are affected in the next survey wave, i.e., three years after the university was awarded excellence status. Consequently, we also include an $Excellent(B)_{jt+1}$ dummy to identify potential long-term effects. We cannot study the isolated labeling effect on students at the University of Freiburg because its excellence status was announced in October 2007, and there was no survey wave during the semester immediately following the announcement. Hence, identification of the $Excellent(B)_{jt}$ effect relies on the three remaining universities of Karlsruhe, Munich, and Dresden , whereas the $Excellent(B)_{jt+1}$ effect also includes the University of Freiburg. Furthermore, we include the following individual-level control variables: age, gender, parents' level of education, field of study, full-time or part-time student status, degree program, school GPA, number of semesters a student has attended university, and whether a student was admitted during the summer term. Dummies for the field of study control for the potentially different experiences of students in different subjects; for example, due to class size. We also control for school GPA because students' ability levels differ between universities, and less academically able students may rate their educational experience worse than their more academically able counterparts. Furthermore, both tuition fees and double cohorts might have an effect on student ratings: The presence of tuition fees might raise students' expectations concerning the quality of education and the intensity of personal support, whereas an instantaneous surge in the number of newly enrolled students due to double high school graduation cohorts might strain a university's facilities and likewise lead to lower

satisfaction ratings. The dummy $TuitionFee(B)_{jt}$ indicates whether a tuition fee was collected at the university, whereas the dummy $DoubleCohort(B)_{jt}$ indicates whether there was a double graduation cohort in the federal state in which the university is located during the time of the survey. Again, we present results both with and without controls for the presence of tuition fees and double high school graduation cohorts and report robust standard errors clustered on university level for all regressions.

### 2.5.2 Results

Tables 2.3 and 2.4 present ordered logit regression results with student ratings of educational quality and job market expectations as the dependent variables, which were standardized to zero mean and unit variance. We differentiate between experience-related items reflecting the educational experience of students and expectations-related items reflecting expected job opportunities and other labor market outcomes. Table 2.3 presents the estimation results for the experience-related items. The dummy variable Excellent(B) identifies students' perceptions of quality of education at universities that were recently announced "excellent". The results reveal that these students rated their university's quality of education significantly better during that semester on dimensions such as quality of curriculum content, quality of teaching, and supervision. Moreover, perceptions of the quality of professional knowledge and practical skills the students acquired while attending university as well as the practical relevance of the material taught was also rated significantly better at recently awarded excellence universities.

Table 2.3: Perceived Quality of Education – Experience-related Items

| Dep. var.: | (1) Content quality | (2) Teaching quality | (3) Supervision | (4) Professional knowledge | (5) Practical skills | (6) Practical relevance |
|---|---|---|---|---|---|---|
| $\text{Excellent}_t$ | 0.247*** | 0.192** | 0.222*** | 0.170** | 0.146** | 0.141** |
| | (0.0749) | (0.0826) | (0.0620) | (0.0705) | (0.0689) | (0.0598) |
| $\text{Excellent}_{t+1}$ | 0.0943 | 0.0160 | 0.110 | 0.0824 | 0.149** | 0.243*** |
| | (0.0709) | (0.0778) | (0.101) | (0.0669) | (0.0700) | (0.0709) |
| GPA Abitur | –0.133*** | –0.0609** | –0.0721*** | –0.346*** | –0.0966*** | –0.0728*** |
| | (0.0254) | (0.0248) | (0.0275) | (0.0201) | (0.0204) | (0.0220) |
| # Semesters | –0.0534*** | –0.0411*** | –0.0110 | 0.0357*** | 0.0507*** | –0.0701*** |
| | (0.0049) | (0.0055) | (0.0081) | (0.0050) | (0.0075) | (0.0046) |
| Tuition fees | 0.0358 | 0.0513 | 0.195*** | 0.0709 | 0.0427 | 0.0243 |
| | (0.0611) | (0.0685) | (0.0461) | (0.0672) | (0.0739) | (0.104) |
| Double cohort | 0.328** | 0.267* | 0.0327 | 0.117 | 0.0621 | 0.138 |
| | (0.150) | (0.144) | (0.138) | (0.0867) | (0.0983) | (0.0967) |
| Observations | 36,865 | 36,847 | 36,833 | 36,881 | 36,861 | 36,694 |
| Pseudo $R^2$ | 0.029 | 0.025 | 0.037 | 0.030 | 0.031 | 0.034 |

*Note*: We regress different survey items (standardized with zero mean and unit variance) on a dummy that identifies ratings of students collected immediately after these universities were awarded excellence status (viz., winter semester 2006/2007 for Munich and Karlsruhe, and winter semester 2012/2013 for Dresden). All the regressions contain wave and university fixed effects. Additionally, all the regressions control for subject of study, degree program, summer-term admissions, part-time study, number of semesters a student has attended university, age, gender, and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Similarly, students at universities that recently received excellence status also significantly adjusted their expectations with respect to their job opportunities. As presented in Table 2.4, the three items show a response of similar magnitude. Since all these items were formulated negatively - for example, by asking about expected difficulties in finding a job - the negative coefficients indicate that the students increased their job expectations. To help with the interpretation of the results, Tables 2.8 and 2.9 in the Appendix report the marginal effects of the $Excellent(B)_{jt}$ dummy at the means of the categories of the respective dependent variable for the models reported in Tables 2.3 and 2.4. The results show that while students whose university was recently labeled excellent are less likely to select a worse category on the questions referring to educational quality or job market expectations, they are more likely to select a better category. For instance, students whose university was recently labeled excellent were 4.3 percentage points more likely than students whose university was not labeled excellent to select response category 6 on a scale

from 1 to 7 (very bad to very good) answering the following question: "What have been your experiences during your studies with respect to the quality of the curriculum's content?" As hypothesized, the students' ratings of both the quality of their education and their job market expectations show significant positive short-term responses to the excellence label. Thus, as students update their beliefs with respect to the quality of their education, they also update their job market expectations. A possible explanation for the fact that the students' ratings of their past educational experiences respond to the excellence label is that the students implicitly benchmark their university against other universities with which they have no or little experience. To corroborate this explanation, we tested whether the students' emotional response to the label - for example, because they identify with their university and feel proud and happy about "being excellent" - might partially drive the positive nature of their ratings and expectations. However, we find that none of the items in the data referring to students' satisfaction unrelated to their belief about their university, such as emotional stress (for example fears and depression) and worries about their personal relationships and financial situation, exhibit any significant response to the award of the excellence label (see Table 2.10 in the Appendix). This finding indicates that students' perceived quality of education response is indeed driven by an update of their beliefs about the relative quality of their institution and not by emotions. The data set also allows us to study whether excellence status has a positive effect on student satisfaction in the long run, i.e., three years after the award when the next wave of data are collected. A possible long-term effect is likely driven not only by the label, but also by the money tied to the award and by the organizational and cultural changes the university underwent due to its new status. As observed in the coefficient of the lead dummy variable $Excellent(B)_{jt+1}$ in Tables 2.3 and 2.4, the evidence that excellence status affects student satisfaction positively in the long run is rather weak. Only the practical skills acquired during one's studies and the practical relevance of one's studies are rated significantly better three years later. The students' responses to all the other experience-related and

expectations-related questions are not significantly more positive three years later, although the universities still enjoy excellence status. However, as shown in Table 2.10 in the Appendix, three years after a university was awarded excellence status, students report more emotional stress from fears and depression, for instance, and seem to worry more about their financial situation. This is an interesting finding the causes of which are worth investigating in further research.

Table 2.4: Perceived Quality of Education – Expectations-related Items

| Dep. var.: | (1) Difficulties in finding a job | (2) Insecure job prospects | (3) Employment worries |
|---|---|---|---|
| Excellent$_t$ | –0.140** | –0.128*** | –0.115*** |
| | (0.0669) | (0.0364) | (0.0386) |
| Excellent$_{t+1}$ | 0.124 | –0.0295 | –0.0744 |
| | (0.116) | (0.0678) | (0.0777) |
| GPA Abitur | 0.189*** | 0.104*** | 0.268*** |
| | (0.0222) | (0.0144) | (0.0185) |
| # Semesters | 0.0446*** | 0.0840*** | 0.0315*** |
| | (0.0051) | (0.0024) | (0.0036) |
| Tuition fees | 0.0445 | 0.00564 | 0.102** |
| | (0.130) | (0.0616) | (0.0437) |
| Double cohort | –0.0328 | –0.0329 | –0.0360 |
| | (0.164) | (0.113) | (0.105) |
| Observations | 33,290 | 36,810 | 36,588 |
| Pseudo $R^2$ | 0.070 | 0.035 | 0.049 |

*Note:* We regress different survey items (standardized with zero mean and unit variance) on a dummy that identifies ratings of students collected immediately after these universities were awarded excellence status (viz., winter semester 2006/2007 for Munich and Karlsruhe, and winter semester 2012/2013 for Dresden). All the regressions contain wave and university fixed effects. Additionally, all the regressions control for subject of study, degree program, summer-term admissions, part-time study, age, gender, and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Our findings for students' perceptions in this section also illustrate an important mechanism underlying the results for admissions in section 4.2. It seems that excellence status causes more students to apply to a university because the award is perceived as a signal of high educational quality and, consequently, better job prospects. As universities have limited capacity and high school grades generally

are the most important selection criterion, "excellent" universities can have more competitive admissions.

## 2.6 Conclusion

Using data from a representative student survey, we investigated whether being successful in the German universities' Excellence Initiative, a competition for research funding, and the accompanying label "university of excellence" allow a university to enroll better students. We found that designated "universities of excellence" recruit students with better high school grades. This effect is statistically significant for three years following the award of excellence status, indicating that the award has a positive effect on student selection for successful universities and increases the ability differences of students at "excellent" and "non-excellent" universities in the medium term. We do not find evidence that the award has a positive effect on the enrollments of successful universities in the longer term. We also investigated an important factor of enrollment decisions: the perceived quality of a university's education. Our findings show that the label "university of excellence" in itself, before any organizational changes due to additional research funds can take effect, has a strongly positive and significant effect on students' satisfaction ratings. Interestingly, this effect is observed even though these ratings refer to past experiences. We hypothesize that this is due to students implicitly comparing their university with other universities with which they have no or little experience when responding to survey items measuring student satisfaction. The award of the label thus causes students to update their beliefs about the relative educational quality of their institution. The fact that following the award of the label, students also adjust their job market expectations but not their satisfaction in areas unrelated to education further supports the hypothesis that the excellence label is perceived as a signal of a university's quality of education vis-à-vis other universities. The actual quality of a university's education, however, does not seem to benefit from the privileged status because ratings of educational quality largely return to previous levels three years

after the award, whereas excellence status persists. By studying a rare and highly publicized event in which information on which universities are considered the best research universities in a country suddenly becomes common knowledge, we provide evidence that there is a clear link between a university's research reputation and student satisfaction ratings. Overall, we find that the research competition resulted not only in stronger competition for (and more inequality of) research funds, which was its declared aim, but also in a more unequal distribution of talented students across universities, an effect that has been found to contribute to increasing wage inequality among graduates (see, e.g., Hoxby and Terry, 1999; Bergh and Fink, 2009). Our results thus shed light on an important implication of competition policies for public universities that has, until now, received little attention in the public debate. So far, however, we can only detect a transitory effect. It remains to be seen whether the effect is reinforced by more universities having their status renewed in further waves of the German Excellence Initiative.

# 2.7 Appendix to Chapter 2

## 2.7.1 Summary Statistics

Table 2.5: Descriptive Statistics

| Variable | Obs. | Mean | Std. dev. | Min. | Max. |
|---|---|---|---|---|---|
| GPA Abitur | 37,642 | 2.198 | 0.633 | 1 | 4 |
| GPA Abitur (stand.) | 37,642 | 0.000 | 1.000 | –1.901 | 2.822 |
| Excellent($A$) | 37,967 | 0.042 | 0.200 | 0 | 1 |
| Excellent($B$) | 37,967 | 0.038 | 0.191 | 0 | 1 |
| Excellent$(B)_{t+1}$ | 37,967 | 0.041 | 0.198 | 0 | 1 |
| *Student Perceptions* | | | | | |
| Content quality | 37,761 | 4.849 | 1.310 | 1 | 7 |
| Professional knowledge | 37,773 | 4.484 | 1.168 | 1 | 7 |
| Practical skills | 37,753 | 2.433 | 1.625 | 1 | 7 |
| Practical relevance | 37,611 | 2.330 | 1.593 | 1 | 7 |
| Teaching quality | 37,741 | 4.324 | 1.351 | 1 | 7 |
| Supervision | 37,727 | 4.074 | 1.496 | 1 | 7 |
| Difficulties to find a job | 34,104 | 2.148 | 0.973 | 1 | 4 |
| Insecure job prospects | 37,721 | 2.594 | 1.915 | 1 | 7 |
| Employment worries | 37,483 | 3.578 | 1.963 | 1 | 7 |
| Stress financial situation | 37,800 | 2.706 | 2.016 | 1 | 7 |
| Emotional stress | 37,746 | 2.218 | 1.895 | 1 | 7 |
| Stress relationship | 37,227 | 1.523 | 1.985 | 1 | 7 |
| *Field of Study* | | | | | |
| Humanities | 37,865 | 0.223 | 0.416 | 0 | 1 |
| Social sciences | 37,865 | 0.138 | 0.345 | 0 | 1 |
| Law | 37,865 | 0.075 | 0.264 | 0 | 1 |
| Economics | 37,865 | 0.126 | 0.332 | 0 | 1 |
| Medicine | 37,865 | 0.102 | 0.302 | 0 | 1 |
| Natural sciences | 37,865 | 0.190 | 0.392 | 0 | 1 |
| Engineering | 37,865 | 0.121 | 0.326 | 0 | 1 |
| Other fields | 37,865 | 0.025 | 0.158 | 0 | 1 |
| *Degree Program* | | | | | |
| Bachelor's | 37,738 | 0.121 | 0.327 | 0 | 1 |
| Master's | 37,738 | 0.038 | 0.192 | 0 | 1 |
| Diplom | 37,738 | 0.384 | 0.486 | 0 | 1 |
| Magister | 37,738 | 0.119 | 0.323 | 0 | 1 |
| State examination | 37,738 | 0.305 | 0.460 | 0 | 1 |
| Other program | 37,738 | 0.020 | 0.140 | 0 | 1 |
| Not defined | 37,738 | 0.008 | 0.091 | 0 | 1 |
| Age | 37,898 | 23.873 | 4.055 | 17 | 83 |
| Female | 37,895 | 0.550 | 0.497 | 0 | 1 |

*Table 2.5* (continued)

| Variable | Obs. | Mean | Std. dev. | Min. | Max. |
|---|---|---|---|---|---|
| Summer admission | 37,967 | 0.144 | 0.351 | 0 | 1 |
| Part-time student | 37,782 | 0.235 | 0.424 | 0 | 1 |
| Semester | 37,967 | 6.637 | 4.316 | 1 | 20 |
| Tuition fees($A$) | 37,967 | 0.166 | 0.372 | 0 | 1 |
| Double cohort($A$) | 37,967 | 0.012 | 0.110 | 0 | 1 |
| Tuition fees($B$) | 37,967 | 0.266 | 0.442 | 0 | 1 |
| Double cohort($B$) | 37,967 | 0.026 | 0.160 | 0 | 1 |
| *Parents' Highest Level of Education* | | | | | |
| Lower secondary (Hauptschule) | 37,904 | 0.090 | 0.286 | 0 | 1 |
| Upper secondary (Realschule) | 37,904 | 0.178 | 0.383 | 0 | 1 |
| High school (Abitur) | 37,904 | 0.143 | 0.350 | 0 | 1 |
| Polytechnic (Fachhochschule) | 37,904 | 0.128 | 0.335 | 0 | 1 |
| University | 37,904 | 0.451 | 0.498 | 0 | 1 |
| Other | 37,904 | 0.010 | 0.100 | 0 | 1 |
| *University* | | | | | |
| TU Berlin | 37,967 | 0.065 | 0.247 | 0 | 1 |
| Bochum | 37,967 | 0.070 | 0.255 | 0 | 1 |
| TU Dresden | 37,967 | 0.096 | 0.295 | 0 | 1 |
| Duisburg-Essen | 37,967 | 0.045 | 0.208 | 0 | 1 |
| Frankfurt | 37,967 | 0.069 | 0.253 | 0 | 1 |
| Freiburg | 37,967 | 0.086 | 0.280 | 0 | 1 |
| Hamburg | 37,967 | 0.088 | 0.283 | 0 | 1 |
| Karlsruhe (KIT) | 37,967 | 0.083 | 0.276 | 0 | 1 |
| Kassel | 37,967 | 0.029 | 0.169 | 0 | 1 |
| Leipzig | 37,967 | 0.094 | 0.292 | 0 | 1 |
| Magdeburg | 37,967 | 0.042 | 0.200 | 0 | 1 |
| LMU Munich | 37,967 | 0.117 | 0.321 | 0 | 1 |
| Oldenburg | 37,967 | 0.020 | 0.141 | 0 | 1 |
| Potsdam | 37,967 | 0.047 | 0.211 | 0 | 1 |
| Rostock | 37,967 | 0.049 | 0.215 | 0 | 1 |

### 2.7.2 Robustness Checks

Table 2.6: Excellence Status and Quality of Admissions (Results When Excluding One Excellence University)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent variable: | Excluding | Excluding | Excluding | Excluding |
| GPA Abitur (standardized) | Dresden | Freiburg | Karlsruhe | Munich |
| Excellent × 1st year | –0.0659 | –0.103*** | –0.0744* | –0.0796* |
| | (0.0391) | (0.0341) | (0.0376) | (0.0449) |
| Excellent × 2nd year | –0.135* | –0.192** | –0.0904** | –0.215** |
| | (0.0727) | (0.0886) | (0.0416) | (0.0981) |
| Excellent × 3rd year | –0.150** | –0.206*** | –0.126* | –0.198* |
| | (0.0643) | (0.0595) | (0.0693) | (0.101) |
| Excellent × 4th year | 0.0694 | 0.103 | 0.0103 | 0.00305 |
| | (0.0683) | (0.0696) | (0.0816) | (0.0930) |
| Excellent × 5th year | 0.101 | 0.0111 | 0.0476 | 0.0952 |
| | (0.0827) | (0.0801) | (0.0908) | (0.0876) |
| Excellent × 6th year | –0.0487 | –0.0557 | –0.0752 | – |
| | (0.0818) | (0.0758) | (0.0757) | – |
| Tuition fees | –0.0299* | –0.0423** | –0.0384** | –0.0427** |
| | (0.0147) | (0.0169) | (0.0177) | (0.0178) |
| Double cohort | –0.0216 | –0.0399 | –0.0113 | –0.0262 |
| | (0.0585) | (0.0632) | (0.0631) | (0.0639) |
| Observations | 35,296 | 35,479 | 35,569 | 34,318 |
| Adjusted $R^2$ | 0.176 | 0.167 | 0.174 | 0.183 |

*Note:* We regress school GPA (standardized with zero mean and unit variance over the whole sample) on six dummies for each year following the award of excellence status. All regressions control for tuition fees and double cohorts, and contain a constant and cohort and university fixed effects. Additionally, all regressions control for field of study, degree program, summer-term admissions, part-time study, age, gender, and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.7: Excellence Status and Quality of Admissions (Results When Excluding Summer Term Admissions)

| Dependent variable: GPA Abitur (standardized) | (1) | (2) | (3) |
|---|---|---|---|
| Excellent | –0.0903* | | 0.103* |
| | (0.0467) | | (0.0502) |
| Excellent × 1st year | | –0.0820* | |
| | | (0.0384) | |
| Excellent × 2nd year | | –0.147* | |
| | | (0.0740) | |
| Excellent × 3rd year | | –0.175** | |
| | | (0.0640) | |
| Excellent × 4th year | | 0.0898 | |
| | | (0.0875) | |
| Excellent × 5th year | | 0.104 | |
| | | (0.0875) | |
| Excellent × 6th year | | –0.0809 | |
| | | (0.0747) | |
| Excellent × Social sciences | | | –0.261*** |
| | | | (0.0373) |
| Excellent × Law | | | –0.245*** |
| | | | (0.0432) |
| Excellent × Economics | | | –0.526*** |
| | | | (0.111) |
| Excellent × Medicine | | | –0.285*** |
| | | | (0.0381) |
| Excellent × Natural sciences | | | –0.157** |
| | | | (0.0684) |
| Excellent × Engineering | | | –0.131 |
| | | | (0.136) |
| Excellent × Other | | | 0.0767 |
| | | | (0.0964) |
| Tuition fees | –0.0777*** | –0.0755*** | –0.0775*** |
| | (0.0218) | (0.0204) | (0.0221) |
| Double cohort | –0.0387 | –0.0770* | –0.0361 |
| | (0.0498) | (0.0419) | (0.0524) |
| Observations | 33,112 | 33,112 | 33,112 |
| Adjusted $R^2$ | 0.171 | 0.172 | 0.172 |

*Note:* We regress school GPA (standardized with zero mean and unit variance over the whole sample) on a dummy that indicates whether a student enrolled in a university after the university was labeled excellent. In column (2) this dummy is separated into six dummies for each year following the award of excellence status. Column (3) contains interaction terms between the excellence dummy and fields of study. Humanities is the reference category. All regressions control for field of study, tuition fees, and double cohorts, and contain a constant and cohort and university fixed effects. Additionally, all regressions control for degree program, part-time study, age, gender and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

### 2.7.3 Further Results

Table 2.8: Marginal Effects of the Excellence Dummy for the Models Reported in Table 2.3

| | (1) Content quality | | (2) Professional knowledge | | (3) Practical skills | | (4) Practical relevance | | (5) Teaching quality | | (6) Supervision | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ |
| 1 | –0.002 | 0.001 | –0.001 | 0.014 | –0.015 | 0.036 | –0.015 | 0.019 | –0.004 | 0.016 | –0.009 | 0.000 |
| 2 | –0.009 | 0.001 | –0.002 | 0.027 | –0.017 | 0.034 | –0.017 | 0.018 | –0.013 | 0.019 | –0.019 | 0.000 |
| 3 | –0.021 | 0.001 | –0.005 | 0.015 | –0.005 | 0.030 | –0.003 | 0.021 | –0.020 | 0.022 | –0.022 | 0.000 |
| 4 | –0.020 | 0.001 | –0.016 | 0.017 | 0.008 | 0.033 | 0.010 | 0.019 | –0.011 | 0.020 | –0.004 | 0.000 |
| 5 | –0.003 | 0.001 | –0.019 | 0.014 | 0.015 | 0.032 | 0.013 | 0.019 | 0.019 | 0.021 | 0.023 | 0.000 |
| 6 | 0.043 | 0.001 | 0.016 | 0.015 | 0.010 | 0.038 | 0.008 | 0.017 | 0.026 | 0.020 | 0.024 | 0.000 |
| 7 | 0.012 | 0.001 | 0.026 | 0.016 | 0.004 | 0.039 | 0.003 | 0.023 | 0.003 | 0.020 | 0.007 | 0.001 |

Table 2.9: Marginal Effects of the Excellence Dummy for the Models Reported in Table 2.4

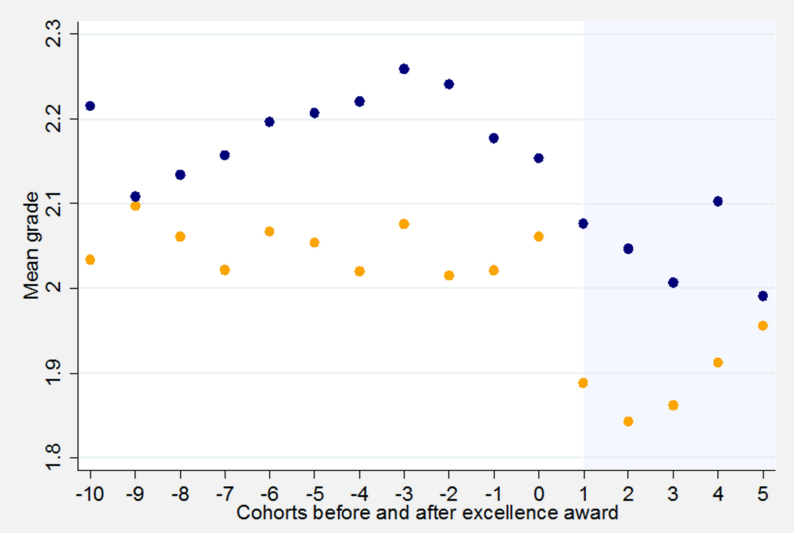
| | (1) Difficulties in finding a job | | (2) Insecure job prospects | | (3) Employment worries | |
|---|---|---|---|---|---|---|
| | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ | $dy/dx$ | $P > \|z\|$ |
| 1 | 0.026 | 0.034 | 0.018 | 0.000 | 0.009 | 0.003 |
| 2 | 0.001 | 0.048 | 0.011 | 0.000 | 0.008 | 0.003 |
| 3 | –0.014 | 0.036 | 0.003 | 0.000 | 0.006 | 0.003 |
| 4 | –0.013 | 0.034 | –0.003 | 0.000 | 0.006 | 0.003 |
| 5 | | | –0.009 | 0.000 | –0.001 | 0.003 |
| 6 | | | –0.011 | 0.000 | –0.010 | 0.003 |
| 7 | | | –0.008 | 0.001 | –0.017 | 0.003 |

Table 2.10: Excellence Status and Emotions

| | (2) Stress financial situation | (3) Emotional stress | (4) Stress relationship |
|---|---|---|---|
| Excellent | –0.0248 | –0.0261 | 0.0374 |
| | (0.105) | (0.0476) | (0.0418) |
| Excellent$_{t+1}$ | 0.155* | 0.155** | 0.0581 |
| | (0.0918) | (0.0606) | (0.0598) |
| GPA Abitur | 0.359*** | 0.155*** | –0.0389** |
| | (0.0219) | (0.0176) | (0.0184) |
| # Semesters | 0.0260*** | 0.0181*** | –0.0191*** |
| | (0.0054) | (0.0039) | (0.0037) |
| Tuition fees | –0.0011 | –0.127** | –0.0039 |
| | (0.146) | (0.0641) | (0.0537) |
| Double cohort | –0.0441 | –0.0215 | –0.103 |
| | (0.115) | (0.0949) | (0.0831) |
| Observations | 36,883 | 36,830 | 36,333 |
| Pseudo $R^2$ | 0.029 | 0.011 | 0.009 |

*Note:* We regress different survey items (standardized with zero mean and unit variance) on a dummy that identifies ratings of students collected immediately after these universities were awarded excellence status (viz., winter semester 2006/2007 for Munich and Karlsruhe and winter semester 2012/2013 for Dresden). All regressions contain wave and university fixed effects. Additionally, all regressions control for subject of study, degree program, summer-term admissions, part-time study, age, gender, and parents' highest level of education. Robust standard errors clustered on university level are reported in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 2.1: Mean Grades by Cohort for Excellence and Non-excellence Universities




*Note:* Blue dots (upper): mean grades by cohort of non-excellence universities; yellow dots (lower): mean grades by cohort of excellence universities.

Table 2.11: Survey Items and Scales

| Item [*variable names in italics*] | Scale |
|---|---|
| What have been your experiences during your studies with respect to ...<br>... the quality of the curriculum's content? [*content quality*]<br>... the way lectures are given? [*teaching quality*]<br>... supervision and counseling by lecturers? [*supervision*] | 1–7<br>(very bad–very good) |
| Please indicate to what extent your studies have promoted your knowledge and skills in the following areas ...<br>... professional knowledge. [*professional knowledge*]<br>... practical skills. [*practical skills*] | 1–7<br>(not at all–very much) |
| How strongly, from your point of view, is your subject of study at your university characterized by ...<br>... good professional preparation/strong practical relevance? [*practical relevance*] | 1–7<br>(not at all–very much) |
| Which of the following options best describes your job prospects after graduation? [*difficulties in finding a job*] | 1–4<br>(hardly any difficulties in finding a job–difficulties in finding any job at all) |
| How much do you personally feel stressed by ...<br>... insecure job prospects? [*insecure job prospects*]<br>... your current financial situation? [*stress financial situation*]<br>... personal problems (e.g., fears, depression)? [*emotional stress*]<br>... the lack of a stable relationship? [*stress relationship*] | 1–7<br>(not at all–very much) |
| What do you think is important for improving your personal situation as a student?<br>... improvement of employment outlook for students of your subject of study [*employment worries*] | 1–7<br>(not at all–very much) |

# Chapter 3

# Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment in Secondary Schools

Co-authored with Valentin Wagner[1]

## 3.1 Introduction

Students and employees are often given feedback about their past performance because it is thought to positively influence their future performance. Feedback[2] has indeed sometimes been found to improve performance (Azmat and Iriberri, 2010; Blanes i Vidal and Nossol, 2011; Tran and Zeckhauser, 2012) and may have advantages over monetary incentives as it can be used when the latter are difficult to implement or not socially accepted. However, feedback is also frequently found to backfire (Barankay, 2012; Ashraf et al., 2014; Azmat et al., 2016; Bradler et al., 2016a) or to be ineffective (Eriksson et al., 2009).[3] Asking which factors are crucial for its success is therefore important.

The influence of a small number of factors on the effectiveness of feedback has al-

[1]My co-author and I contributed equally to the design and implementation of the study, to the data analysis, and to writing the paper.

[2]Economists have investigated different kinds of feedback, such as process feedback (by allowing subjects to observe the behaviors of other people performing the same task, see e.g. Falk and Ichino, 2006; Mas and Moretti, 2009) or outcome feedback (by providing a quantitative measures of past performance such as a test score or rank, see e.g. Tran and Zeckhauser, 2012; Azmat et al., 2016). We will focus on outcome feedback in this study.

[3]See also Kluger and DeNisi (1998) for evidence from the psychological literature.

ready been investigated. For example, it has been found that the effects of feedback depend on whether a pay-for-performance or a flat incentive scheme is present (Azmat and Iriberri, 2016), or whether the information provided is sufficiently precise (Hannan et al., 2008). Furthermore, relative feedback, such as a performance rank, has been found to be more effective than performance information referring to an absolute standard, such as test score (Azmat and Iriberri, 2010). There are mixed findings about whether giving rank information in public or private is more effective (Tran and Zeckhauser, 2012; Hannan et al., 2013; Tafkov, 2013; Ashraf et al., 2014; Gill et al., 2016).[4] Besides these findings, the question of what makes feedback effective has received rather little attention, leaving many aspects that could be relevant for its success as a motivational tool unstudied. This paper begins to fill this gap by studying whether the timing and the reference frame of feedback influence its effectiveness.

In this paper, we study a field experiment in secondary schools in which we exogenously vary *whether* students receive private rank feedback, *when* they receive it and *what* its standard of comparison (reference frame) is. Students aged around 11-12 years in secondary school classes received private written feedback from their teachers and it either contained (i) information about their absolute rank in the last math exam (level feedback), (ii) information about their change in ranks between the two previous math exams (change feedback), or (iii) no information. Students received the feedback either (a) 1-3 days or (b) immediately before the last math exam of the school year. As mathematics is a core subject of the curriculum and students write six exams in this subject during a school year, their performance in the final exam influences whether they will be allowed to stay in their current educational track and to progress to the next grade.

We chose to provide students with relative performance information as people are strongly motivated by it, even in the absence of any tangible benefits (Charness and Rabin, 2002; Azmat and Iriberri, 2010; Kuziemko et al., 2014; Gill et al., 2016).

[4]See also **?** for a summary of the findings in the tournament literature.

If individuals are on average overconfident with respect to their performance level, as has often been found in other settings (Krueger and Mueller, 2002; Hoelzl and Rustichini, 2005; Malmendier and Tate, 2005; Park and Santos-Pinto, 2010), level feedback should make them less confident in already having done "enough", which has been found to positively influence performance (Azmat et al., 2016). However, due to strong complementarities of skill formation at different stages of the education production function, the differences in academic skills increase over time (Cunha and Heckman, 2007) such that large ability differences can often already be found at the ages we study. We know from the literature on tournaments (Gürtler and Harbring, 2010) that revealing information about performance levels may reduce motivation when there is large heterogeneity among them. Feedback that compares students not in terms of their levels but in terms of their changes in performance may alleviate this problem while possibly maintaining the motivational effects of social comparison. More importantly, feedback about how one's performance has changed in the past may also help to promote the belief that skills can be developed by exerting effort, also called a "growth mindset" in the psychological literature (see O'Rourke et al., 2014; Paunesku et al., 2015, which is closely related to the concept of "grit", recently investigated by Alan et al., 2016).

Timing is potentially crucial for the effects of feedback because outcomes in the workplace or educational settings may be influenced by effort exerted at different times – on preparation and on the task itself (cf. Levitt et al., 2016; Wagner, 2016). While earlier feedback may influence preparation effort, possibly by countering students' tendency to procrastinate and start preparations too late (Steel, 2007), feedback given more immediately before a task may potentially have a stronger effect on effort at the task itself due to people's tendency to place a greater weight on more recent information (Hogarth and Einhorn, 1992). Furthermore, timing of feedback may also matter if it influences both expectations and emotions (Loewenstein, 2000; Lane et al., 2005; Kräkel, 2008; Bradler et al., 2016b) and the latter have stronger effects on motivation in the short run than in the long run (Lempert

and Phelps, 2014). For example, someone who learns that his past performance is worse than expected may realize that he has to work harder to attain his desired outcome. However, having this overconfidence corrected may involve (temporary) negative emotions that decrease the enjoyment of a task or distract from it (Benabou and Tirole, 2016) and may thus decrease performance in the short run.[5]

We find that feedback increases subsequent performance when given a few days before the exam and that change and level feedback are equally effective. In classes with early feedback, students receiving feedback about their rank level significantly increase their performance by 0.2 grade points (3.9 percentage points) compared to students receiving no feedback, while students receiving feedback about rank changes significantly increase their performance by 0.3 grade points (3.8 percentage points). We find it to be particularly beneficial to inform students who became worse about their negative change in performance a few days before the exam as this significantly improves these students' outcomes by 0.6 grade points (8.1 percentage points). In contrast, any feedback given to students immediately before the exam tends to lower subsequent performance but the overall effects are not significant. However, informing students who became worse about their negative change in performance immediately before the exam decreases these students' outcomes significantly by 0.3 grade points (but the effects on these students exam scores are not significant).

To shed light on the mechanisms that drive the effects of feedback on performance, we elicit students' belief in the effectiveness of their effort and their emotions captured by their state self-esteem. Our findings show that change feedback, but not level feedback, has a weakly significant positive effect on students' belief that they can affect their outcomes by exerting effort. We also find that both types of feedback tend to have a negative effect on students' state self-esteem. Subgroup analyses reveal that the positive response to early feedback is mostly driven by boys and that boys' self-esteem is strongly reduced by feedback while we do not find

[5]The importance of timing is also supported by the dual-process theory that has found its way into behavioral economic models in recent years (Loewenstein, 2000; Alos-Ferrer and Strack, 2014): People's immediate "hot state" response to information likely differs from their longer term "cold state" response.

negative effects of feedback on the self-esteem of girls. Furthermore, we do not find significant heterogeneity in the effects of feedback by confidence in mathematics abilities, locus of control, or preference for competition.

To our knowledge, this is the first study identifying causal effects of timing of feedback and the first to compare the causal effects of two generic types of feedback (about relative levels and relative changes of performance). As far as we know, it is also the first study to use experimental variation to cleanly identify the causal effects of feedback information in schools. Our results are not only relevant for educators but the general findings extend to other settings where feedback is given with the intention to increase motivation, such as the workplace or the healthcare system.

The paper is organized as follows. The next section gives a brief overview of the related literature. In section 3.3 we motivate the treatment variation and report the results of a survey conducted prior to the experiment in which we test whether students of our target age group understand and how they perceive the two types of feedback. Section 3.4 describes our experimental procedure. Section 3.5 presents the results and investigates potential behavioral mechanisms driving these results. Section 3.6 concludes.

## 3.2 Related Literature

Besides screening for talent[6], economists traditionally focus on the introduction of incentives to raise performance. In recent years, field experiments on monetary (Angrist and Lavy, 2009; Kremer et al., 2009; Fryer, 2011; Bettinger, 2012; Fryer et al., 2012; Levitt et al., 2016) and non-monetary (Jalava et al., 2015; Wagner and Riener, 2015; Levitt et al., 2016) incentives for teachers and/or students have produced mixed results.[7]

Few studies so far have investigated at the effects of feedback in the context

[6]Surprisingly little of the large heterogeneity of teacher effectiveness can be explained by observable teacher characteristics (Hanushek and Rivkin, 2006), which makes it difficult to improve educational outcomes by screening for good teachers.

[7]Damgaard and Nielsen (2017) recently review the use of behaviorally motivated interventions in education.

of education and among those we are aware of, all but one (Azmat and Iriberri, 2010) have relied on university student samples. Tran and Zeckhauser (2012) provide Vietnamese students participating in an experiment involving an English test either with private feedback (by phone) or private plus public feedback (postings on the university's noticeboard and website) about their ranking in in-course mock exams. Overall, the authors find a positive effect of feedback on the final English test and that private plus public feedback tends to outperform private feedback alone. This difference, however, was only marginally significant.[8] A more recent study by Bandiera et al. (2015) exploits data of a natural experiment in the UK where some university students were provided with private, absolute feedback on their past exam performance and others were not. Feedback on exam performance improved future performance mostly for more able students and for students who initially had less information about the academic environment. Azmat et al. (2016) provide college students with feedback on their position in the grade distribution every six months over a period of three years. They find that students who received feedback suffered a decrease in their performance relative to a control group. This effect is driven by students who underestimated their relative performance in the absence of feedback.

While these studies analyze the effect of feedback on performance among university students, we are aware of only one study on school aged children which exploits data from a *natural* field experiment and there is – to our knowledge – no *randomized controlled* field experiment on the effectiveness of performance feedback on educational outcomes of children. Azmat and Iriberri (2010) study the motivational effect of relative performance feedback among high school students in Spain (aged 14 - 18) in a natural field experiment. For one school year, a high school in the Basque Country adopted a new system of producing report cards providing students with information on whether they were performing above or below the class average as well as the distance from this average. Before and after this change, report cards informed students only about their own grade point average. The new

[8] In contrast to Tran and Zeckhauser (2012), Ashraf et al. (2014) find that private plus public feedback reduces performance of health workers in Zambia in a nationwide training program.

relative performance feedback had positive effects and increased students' grades by 5 %. However, the effect disappeared as soon as the information was removed.

The paper by Azmat and Iriberri (2010) is the one most similar to ours with respect to the population studied. With respect to the dimensions of feedback – timing and social reference frame – manipulated in our design, we are not aware of any similar studies.

## 3.3 Motivation and Pre-test of Treatments

### 3.3.1 Motivation of Treatments

We varied both the type as well as the timing of feedback as we expected both dimensions to matter for how feedback affects behavior. We expected that level feedback influences students' empirical beliefs in different ways than change feedback. Building on a model by Fischer and Sliwka (Chapter 5) we expected that two types of beliefs matter for how much effort a student invests in the exam, (i) confidence in her past level of math performance and (ii) confidence in the effectiveness of her effort (i.e. her ability to improve her math performance). Assuming that students at different parts of the ability distribution each strive for exam outcomes within their reach, the model predicts that increasing a student's confidence in her past level of math performance decreases the necessity to invest additional effort in exam preparation to reach the desired outcome in the next exam. Furthermore, according to this model, confidence in the effectiveness of effort reduces a person's perceived effort costs. Thus, raising confidence in the effectiveness of effort increases effort. Fischer and Sliwka (Chapter 5) find that people's effort in a lab experiment responds as predicted by their model.

In our classroom setting, the effect of feedback about one's level of past performance depends on whether a person ex-ante is overconfident or underconfident with respect to her level of past performance. If she is overconfident, learning about the true level of past performance is disappointing and thus will lower her confidence

in her level of performance (and increase the perceived necessity of effort), if she is overconfident learning the same information will be positively surprising and raise her confidence in her level of performance (and decrease the perceived necessity of effort). Likewise, the effect of feedback about one's changes in performance depends on how it affects a person's beliefs. The psychological literature on the so called "growth mindset" (O'Rourke et al., 2014; Paunesku et al., 2015) argues that making changes in past performance salient strengthens confidence that one's outcomes can be influenced by one's effort, i.e. confidence in the effectiveness of effort (which increases effort).

In recent years, economists have started to consider the role of emotions in decision making. Possibly emotions besides information processing may mediate the effort response to feedback. Disappointing feedback likely worsens a person's emotional state, which may decrease the enjoyment of a given task and effort (Lane et al., 2005; Benabou and Tirole, 2016). Emotions are generally considered to be short lived (Lempert and Phelps, 2014). For this reason the short-run response to feedback may differ from the longer-run response(Loewenstein, 2000; Alos-Ferrer and Strack, 2014), which implies that the timing of feedback may be relevant for the observed response. While the longer-run behavior may be driven by the rational response to new performance information, the short-run response may be driven by a combination of rational and emotional response. For example, in the short run disappointing performance feedback raises the necessity to exert more effort (strengthening extrinsic motivation) but may at the same time worsen the emotional state (weakening intrinsic motivation). Thus while disappointing information likely increases effort in the longer run, the emotional response attenuates the incentive effect and may even dominate it in the short run.

### 3.3.2 Pre-test of Treatments

When teachers return the graded exams to students, they often provide them with a statistic about the frequency of grades in their class. Students therefore have some

imprecise information about how their performance compares to the performance of other students. Students in our sample are quite young and in order to test whether they understand our feedback (to disentangle lack of understanding and ineffectiveness of the information) and how they interpret it (to enable us to interpret possible effects), we conducted a survey in 6 classes in 4 schools with a total of 151 students of the same age group as our experimental sample before implementing the field experiment. This was a convenience sample gathered through personal contacts.

The survey consisted of a two-page questionnaire. On the front of the page students saw a feedback note of a fictitious student named "Paul" and were asked to imagine themselves in his position. The feedback note contained either level or change feedback, and both of them were varied (good ranks to bad ranks, positive and negative change in ranks). On the back of the page, students had to shortly summarize the information on the front of the page and answer a quiz to test whether they understood it correctly. They were also asked to give their guess of how Paul feels ("very good" to "very bad") after having read the feedback note and of how highly motivated ("not at all" to "very strongly") Paul will be to exert effort in the next exam. We also asked students whether they knew the size of their class, which is crucial for correctly interpreting rank feedback.

Most students correctly understand the feedback notes. 85.56% of the students could correctly calculate by how much Paul's rank changed and 94.74% could correctly determine the position of Paul's rank when given level feedback. Moreover, 86.09% of students know the exact size of their class. The mean responses to the questions concerning Paul's emotions and motivation are presented in Figure 3.1 in Appendix 3.7.1.[9] Students believe that bad feedback (negative change in ranks or rank level below median) makes a student feel worse than good feedback but that the student's motivation to exert effort is quite high (above 3 on a 5 point scale)

[9]The results indicate that students believe that Paul would be more motivated when receiving change feedback than when receiving level feedback while they do not indicate that the two feedback types affect emotions differently. Note that the difference in reported motivation between the change feedback and the level feedback may be driven by the chosen ranks.

and approximately the same with negative and positive feedback.

Overall, the results of the pre-experimental survey indicate that most students of our target age group correctly understand the information contained in two types of feedback and that they perceive their content as affecting emotions but do not believe that more negative feedback will generally be less motivating than more positive feedback.

## 3.4 Experimental Intervention

The experiment was conducted in 19 classes (grades 5 and 6) in 7 secondary schools in the German cities of Bonn, Cologne, and Düsseldorf and was approved by the ethics committee of the University of Düsseldorf. 352 students received parents' consent (73.9% per class) and participated in the experiment during May and June 2016.[10] Researchers were never present in the classroom to maintain a natural examination situation and the feedback was given to students by their math teacher to maximize its credibility.[11] To train teachers how to conduct the experiment, we visited the schools in the run-up of the experiment. During this meeting, the intervention was explained and teachers' questions were answered. We sent teachers two envelopes with material needed to run the experiment. A first envelope contained written teacher instructions outlining the time schedule and steps of the intervention, consent forms to be signed by parents and templates for providing results of the fourth and the fifth math exams of the school year, consisting of the classes' grades and points in each of the two exams and the maximum number of points reachable. For those students whose parents consented, teachers provided us with names, enabling us to print personalized feedback notes by calculating students' ranks in the last math exam and their change in ranks from the second last to the

[10] We contacted 142 secondary schools in the federal state of North Rhine-Westphalia (NRW) using a list of schools that is publicly available from the Ministry of Education of NRW. 23% of the schools responded and 39% (13 out 33) of these schools were generally interested in participating. After further consultation with schools, 7 schools finally participated.

[11] The credibility of the source has a substantial effect on how feedback is interpreted. Ilgen et al. (1979) identified two components of source credibility: expertise and trustworthiness.

last math exam. A second envelope was sent to schools a few days before the third exam. It contained the personalized feedback notes, which were sheets of paper that were folded and had the name of the student it referred to clearly written on its outside. The envelope also contained a result template for the third exam and student questionnaires.

**Treatments**

We want to test how relative performance feedback affects a student's performance in a high-stakes math exam. As described above, relative feedback has often proven effective in raising performance but has also been found to backfire and there is little evidence on the effects of feedback in schools. Rank feedback also seems promising in light of recent findings that a student's rank within their class or cohort affects later achievement independently of underlying ability (Murphy and Weinhardt, 2014; Elsner and Isphording, 2017).[12] Based on a 2 X 3 design, we vary both the *timing* of feedback and the *reference frame* of feedback independently. We are not aware of any studies that have looked at the effect of timing, although it is potentially very important because test outcomes are influenceable both by learning and test taking effort, exerted at different times. Furthermore, feedback can be given in terms of individual levels of performance (rank in last test) and in terms of changes of performance (change in rank between second last and last math test). While all prior studies on rank feedback we are aware of have used levels, the tournament literature points towards this being harmful in settings where ability differences are large (Gürtler and Harbring, 2010), such as in many classrooms. There is also evidence from the psychological literature that promoting the belief that own skills are changeable improves a student's motivation (O'Rourke et al., 2014; Paunesku et al., 2015).

[12]Murphy and Weinhardt (2014) find that students with a one standard deviation higher rank in primary school will score 0.08 standard deviations better at age 14 and Elsner and Isphording (2017) find that high school students with a higher rank have higher expectations about their future career outcomes, are more optimistic and self-confident and, indeed, have a higher likelihood of going to college.

The timing of feedback was randomized on class level. Students either received feedback 1-3 days before the exam (Early Timing) or immediately before the exam sheets were handed out (Late Timing). The reference frame of feedback was randomized at student level. Within the same class, students with parents' permission to participate received personalized written feedback about their rank in the last math exam (Level Frame), about their change in rank between the second last and the last math exam (Change Frame), or a personalized note that only wished them good luck (Control). In all treatments, teachers gave a folded feedback note to each student that had the student's name written on its outside. To personalize the feedback, the note addressed the student by their first name and was signed with the teacher's name (see Appendix 3.7.2 for English translations of the exact wording and layout of the notes). While students in Control received no information about their past performance, in Change Frame, students received information about their change in rank between the two previous exams but no information on their absolute ranks in these tests ("*I compared the points of each student in the class in the last two exams. Relative to your classmates, you improved (worsened) your performance in the last math exam by XX places.*"). Students in Level Frame were informed about their relative rank in the last exam but received no information on their performance in the second last exam or about how their performance changed ("*I looked at the points of each student in the class in the last exam. Relative to your classmates you achieved, with your performance in the last math exam, the XX th place.*"). As students had received their grades in the last two exams after the teachers had graded them (i.e. approximately 2 and 4 months before the last exam, respectively), the feedback information served as a reminder that contained more detailed information about different aspects of their relative performance.

In Early Timing, students had to fill in a questionnaire immediately after receiving the feedback notes, while in Late Timing students had to fill in a questionnaire immediately after completing the exam. Due to time constraints, in Late

Timing, the questionnaire was shorter and did not include all scales included in the Early Timing questionnaire. The questionnaire elicited effort-effectiveness beliefs, preference for competition, character traits, and demographic information. It enables us to study whether the feedback possibly affects test outcomes by changing beliefs about how easily outcomes can be affected by effort. Furthermore, feedback can possibly have heterogeneous effects on students with different gender (Buser and Yuan, 2016) and character traits (Ilgen et al., 1979; Lam and Schaubroeck, 2000; Noe, 2000; Fedor et al., 2001). The questionnaire enables us to explore these possible differences. Questions on character traits are based on validated questionnaires and measured locus of control ( adapted from PISA, based on Rotter, 1966), confidence in math ability (adapted from PISA, based on Bandura 1986) and self-esteem (German version by von Collani and Herzberg (2003)of the Rosenberg self-esteem scale (Rosenberg, 1965), slightly adapted for age). Additionally we elicited preference for competition with questions adapted from PISA.[13]

After students filled in the questionnaires, teachers collected them, while students were required to crumble the feedback notes and throw them in a garbage bin.[14] Upon sending the results of the final exam as well as the filled-in questionnaires, teachers were asked to fill in a short online survey.

## 3.5 Results

This section presents the results and is organized as follows: First, we describe our randomization strategy and discuss concerns about non-random self-selection into treatment groups. Thereafter, we present our data and descriptive statistics before analyzing the impact of feedback on students' performance. We first investigate the effects of timing and then of the reference frame of feedback.

[13]For the measures adapted from the PISA studies also see Marsh et al. (2006).

[14]This was to prevent the feedback notes from being shown to other students (with Early Timing) and from teachers finding them in the exam booklets when they graded the exams (with Late Timing).

### 3.5.1 Randomization and Self-selection

Blocked on school level, classes were randomized either into the Early Timing treatment or the Late Timing treatment. With respect to these class-level treatments non-random self-selection was possible as parents learned whether feedback would be given 1-3 days before the exam or immediately before the exam. This was necessary to receive parents' fully informed consent. Within classes, students were then randomized into the Control group, Change Frame treatment or Level Frame treatment. Parents did not learn to which of the three treatments their child was assigned as randomization into student-level treatments took place only after we obtained parents' consent and students only learned it when they received their feedback notes. Hence, non-random self-selection into the student-level treatments was not possible.

Overall, randomization for both class-level and student-level treatments was successful as no significant differences between treatments are found in any relevant dimensions (prior test scores and grades, gender, student demographics). In the following we will discuss the randomization checks in detail. Table 3.6 in Appendix 3.7.3 reports differences between Early Timing and Late Timing. Student and teacher observables do not differ significantly between these class-level treatments, except with respect to the share of students per class who participated and teacher experience. Fewer students per class participate in Early Timing as compared to Late Timing and teachers in Early Timing are more experienced than teachers in Late Timing.

Surprisingly, the share of participants turned out to be significantly lower in the Early Timing treatment as compared to the Late Timing treatment. We expected the opposite as parents might be concerned about larger negative (emotional) effects on exam outcomes of their children when feedback is given shortly before the exam.[15] This could be an indication that parents were not concerned

[15]Overall, 26.1% students did not get their parents' consent to participate in the experiment (22.5% in the Late Timing treatment and 29.7% in the Early Timing treatment). In 16 out of 19 classes, more than 50% of the students within the class participated.

about the timing of the feedback and that the difference in participation rates is just a coincidence, in particular because all relevant characteristics are balanced. Furthermore, as the treatment groups are balanced on student characteristics, we do not expect teacher experience to influence our results. Teacher characteristics, such as education or experience, do not explain much of the variation in educational outcomes (Rivkin et al., 2005). Moreover, our analysis controls for teacher grading by accounting for prior test scores and by standardizing test scores on class level.

Randomization checks for student-level treatments (Change Frame, Level Frame, Control) can be found in Table 3.7 and Table 3.8 in Appendix 3.7.3. As mentioned above, self-selection into these treatments was not possible, as students had no information on assignment prior to the intervention, and student observables in the student-level treatments are not significantly different from each other.

To summarize, a lower proportion of students participate in the Early Timing treatment. However, student characteristics and prior performance measures do not differ significantly between the class-level and the student-level treatments.

### 3.5.2 Data and Descriptive Statistics

Our data consist of pre and post intervention performance measures provided by the teachers as well as information from student questionnaires. Importantly, we have detailed information on students' past performance as we know students' grades and points in the two last exams before the interventions as well as the maximum score possible in the exams. This information can be treated as exogenous, and may be use in the analysis to control for heterogeneity in ability, because students wrote the exams several months before teachers learned about the study. Students are on average 11.60 years old and have 1.33 siblings. 46.42% of the students are female and 38.04% of students have a non-German first and family name, hinting at a recent migration experience in their family. The average grade in exam 1 is 2.74 and 2.59 in exam 2 on a scale from 1 to 6, where 1 is the highest and 6 is the lowest

grade.[16] Table 3.1 summarizes the feedback students received by treatment and reveals that the range and standard deviation of feedback received in the Change Frame and Level Frame treatments are of similar magnitude. Figures 3.8 and 3.9 in Appendix 3.7.5 show the distribution of given feedback pooled over class-level treatments.

Table 3.1: Descriptive statistics of provided feedback

| | | Obs. | Mean | Std. Dev | Min. | Max. |
|---|---|---|---|---|---|---|
| Change Frame | Early Timing | 59 | 0.763 | 8.052 | -21 | +21 |
| | Late Timing | 57 | 0.842 | 8.239 | -19 | +19 |
| Level Frame | Early Timing | 64 | 13.922 | 8.407 | 1 | 30 |
| | Late Timing | 60 | 13.233 | 8.208 | 1 | 30 |
| Control | Early Timing | 55 | - | - | - | - |
| | Late Timing | 55 | - | - | - | - |

*Note:* This table presents descriptive statistics of the feedback given to students by class-level and student-level treatments.

### 3.5.3 Effects of Feedback on Performance

In the following we investigate the effects of our intervention. We first analyze the effect of *timing* of feedback (Early Timing versus Late Timing) on performance, which was randomized at the class level. Then we will analyze the overall effect of the *reference frame* of feedback (Change Frame versus Level Frame versus Control), which was randomized at the student level. The following tables present results from linear regressions (OLS) that include prior performance as linear control variables and student characteristics as dummy variables, as well as a constant. Furthermore, all regressions contain class fixed effects. The advantage of including class fixed effects is that we can control for heterogeneity of the class environments and the identified effects of feedback are based on comparing students within the same class. For all presented results, the reported standard errors are clustered at the class level and corrected using bias-reduced linearization (Bell and McCaffrey,

[16] Approximate translation of German grades to American grades: 1.0 to 1.3 =A;>1.3 to 2.3=B; >2.3 to 3.3=C; >3.3 to 4.0= D; >4.0=F

2002; Angrist and Pischke, 2008; Cameron et al., 2008; Cameron and Miller, 2015) to allow for cluster-robust inference with a small number of clusters.

First, we study the effect of *timing* of feedback on performance to learn whether students receiving the intervention 1-3 days before the exam had different outcomes than students receiving the intervention immediately before the exam. Then, we will look at the Early Timing and the Late Timing groups separately to study the effect of *reference frame* of feedback. This will allow us to explain whether a possible difference between the Early Timing and the Late Timing groups is driven by the effects of the Change Frame, or the Level Frame, or by both.

**The Role of Timing of Feedback**

We will investigate how timing affects the effectiveness of feedback by comparing students who receive feedback with students who did not receive any feedback *within* both Early Timing and Late Timing classes. To investigate whether there were spillover effects of our intervention on the control group in Early Timing classes (which was not possible in Late Timing classes) we will then compare the results of the control groups of Early Timing and Late Timing classes.

To analyze whether receiving feedback was beneficial at either or both points in time, we estimate the following OLS model *separately* for classes who had the intervention early and classes who had the intervention late:

$$PointsTest3_i\ (GradeTest3)_i = \alpha + \beta\, Feedback_i + \gamma\, PointsTest1_i + \delta\, PointsTest2_i + \\ + \eta\, Covariates_i + \theta\, Class_j + \varepsilon_{ij} \tag{3.1}$$

$PointsTest3_i$ are the percentage points in the final math exam of student $i$, $PointsTest1_i$ and $PointsTest2_i$ are the percentage points in the second last and the last exam of student $i$, $Covariates_i$ is a vector of characteristics of student $i$: student $i$'s gender, whether student $i$ has a non-German name (to capture migration

background), whether student $i$ has siblings, and whether student $i$ has his own room at home. $Feedback_i$ indicates whether student $i$ received feedback or not while $Class_j$ controls for class fixed effects such that $Feedback_i$ identifies the effect of feedback by comparing the results of students who received feedback with those of their classmates who did not. $\varepsilon_{ij}$ is a stochastic i.i.d. error term. While the number of points attained by a student in the final exam captures his level of math knowledge, which is the socially relevant outcome, the student himself might only care about his grade. For this reason, we re-estimate the model with students' grades in the final exam ($GradeTest3_i$) as dependent variable to investigate whether the intervention affected the outcome that may be most relevant for the student.

Table 3.2: Effects of Feedback in Early Timing and Late Timing

| | Dep. Var.: Points in Exam 3 | | Dep. Var.: Grade in Exam 3 | |
|---|---|---|---|---|
| | If Early Timing | If Late Timing | If Early Timing | If Late Timing |
| Feedback | 0.038** | -0.012 | -0.238** | 0.142 |
| | (0.017) | (0.016) | (0.093) | (0.096) |
| Points Exam 1 | 0.358*** | 0.127 | -2.588*** | -1.529*** |
| | (0.046) | (0.121) | (0.438) | (0.574) |
| Points Exam 2 | 0.296*** | 0.436*** | -1.980*** | -2.814*** |
| | (0.068) | (0.110) | (0.516) | (0.499) |
| Female | 0.005 | -0.040 | -0.019 | 0.154 |
| | (0.028) | (0.030) | (0.181) | (0.156) |
| ClassFE | Yes | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes | Yes |
| $N$ | 160 | 159 | 160 | 159 |
| adj. $R^2$ | 0.520 | 0.362 | 0.547 | 0.396 |

*Note:* This table presents the effect of feedback timing on performance in the last exam using a linear regression model including class fixed effects. Columns 1 and 3 present results for classes in which some students received feedback 1-3 days before the exam while columns 2 and 4 present results for classes in which some students received feedback immediately before the exam. The dependent variable in columns 1 and 2 is percentage points in exam 3. The dependent variable in columns 3 and 4 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 10 in model 1 and 9 in model 2. * p<0.10, ** p<0.05, *** p<0.01.

The first and the third column of Table 3.2 show that students who received feedback 1-3 days prior to the exam had on average 3.8 percentage points more and about 0.2 better grades, on a scale from 1.0 (best grade) to 6.0 (worst grade), than students who did not receive any feedback. These effects are both significant at the 5% level. However, when looking at the second and the fourth column, we can see that students who received feedback immediately before the exam did not have significantly different results from students who did not receive any feedback. Results when excluding class fixed effects, prior performance measures, and student characteristics can be found in Table 3.10 in Appendix 3.7.5.

**Possible spillover effects** Note that the above analyses identify the effects of timing on performance by comparing students who received feedback in the EARLY TIMING classes with their classmates who did not receive any feedback and by comparing students who received feedback in the LATE TIMING classes with their classmates who did not receive any feedback. No spillover effects of feedback are possible in LATE TIMING classes as students could not find out anything about the feedback other students received (all students were already seated separately to write the exam and received sheets formatted in the same way). However, the positive effect of feedback in EARLY TIMING classes could possibly be driven by spillover effects of our intervention on students who did not receive any feedback. For example, students who found out after our intervention but before the exam that their classmates received feedback while they did not could have been discouraged, leading them to perform worse in the exam compared to a situation where their classmates were not treated. This would cause the positive effect of feedback to be overestimated. Alternatively, the spillover effects could go in the other direction and students who did not receive any feedback in the early treatment could, by interacting with those who did receive feedback, become more motivated and perform better in the exam. This would cause us to underestimate the benefits of feedback in the early treatment. To address the question of whether there were spillover effects in EARLY TIMING classes, we compare the results of students in the control groups of EARLY TIMING (where

spillover effects were possible) and LATE TIMING classes (where spillover effects were not possible).

As can be seen in Table 3.9 in Appendix 3.7.5, there are no significant differences between the control groups of classes who received the intervention 1-3 days before the exam and classes who received the intervention immediately before the exam in terms of points or grades in the final exam. Interestingly, the results indicate that students in the control group in classes where spillover effects were possible (EARLY TIMING) tended to have better outcomes than their counterparts in classes where not spillover effects were possible (LATE TIMING). We infer from this that, if anything, spillover effects of our intervention on the control group were positive and that the positive effects of early feedback reported in Table 3.2 are lower bound estimates, i.e. we tend to underestimate these effect.

In the next section we will analyze whether the different reference frames of feedback matter for its effectiveness.

### The Role of Reference Frame of Feedback

In order to investigate the role of reference frame of feedback, we will estimate the following model:

$$\begin{aligned} PointsTest3_i\ (GradeTest3)_i = \alpha + \beta\, ChangeFeedback_i + \gamma\, LevelFeedback_i + \\ \delta\, PointsTest1_i + \zeta\, PointsTest2_i + \eta\, Covariates_i + \theta\, Class_j + \varepsilon_{ij} \end{aligned} \tag{3.2}$$

$PointsTest3_i$ are the percentage points and $GradeTest3_i$ is the grade student $i$ in the final math exam.$PointsTest1_i$ and $PointsTest2_i$ are the percentage points in the second last and the last exam of student $i$, $Covariates_i$ is the same vector of characteristics of student $i$ as in equation 3.1. $Class_j$ controls for class fixed effects and $\varepsilon_{ij}$ is a stochastic i.i.d. error term.

We will analyze this model separately for classes who had the intervention 1-3 days before and classes who had the intervention immediately before the exam in order to investigate why students seem to benefit from receiving feedback 1-3 days but not from receiving feedback immediately before the exam.

**Change and level feedback given early** Table 3.3 presents the results with respect to the reference frame of feedback for classes that were treated 1-3 days before the exam. As can be seen in the first and fourth column ("All"), when given early, both types of feedback lead to higher exam scores and better grades than those of students who did not receive any feedback. Students who received change and students who received level feedback have 3.8 and 3.9 percentage points higher outcomes (0.2 and 0.3 better grades) , respectively, than students in the control group. These effects are significant at the 10% and the 5% level (at the 10% and the 1% level).

Table 3.3: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | 0.038* | 0.002 | 0.081*** | -0.220* | 0.048 | -0.588*** |
| | (0.022) | (0.051) | (0.027) | (0.130) | (0.353) | (0.171) |
| Level Frame | 0.039** | 0.026 | 0.053** | -0.254*** | -0.181 | -0.386** |
| | (0.016) | (0.037) | (0.025) | (0.092) | (0.237) | (0.164) |
| Points Exam 1 | 0.358*** | 0.318** | 0.473*** | -2.581*** | -2.367** | -3.655*** |
| | (0.046) | (0.149) | (0.127) | (0.440) | (1.096) | (0.939) |
| Points Exam 2 | 0.297*** | 0.350*** | 0.161 | -1.988*** | -2.407** | -0.543 |
| | (0.067) | (0.128) | (0.121) | (0.511) | (0.967) | (0.980) |
| Female | 0.005 | -0.006 | 0.020 | -0.017 | -0.005 | -0.072 |
| | (0.029) | (0.051) | (0.022) | (0.184) | (0.325) | (0.125) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 160 | 87 | 73 | 160 | 87 | 73 |
| adj. $R^2$ | 0.517 | 0.426 | 0.611 | 0.544 | 0.481 | 0.632 |

*Note:* This table presents the effect of change frame and level frame feedback when given 1-3 days in advance using a linear regression model including class fixed effects. Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Since the effects of change feedback may depend on whether it reported a positive or a negative change, columns 2 and 3 and columns 5 and 6 further investigate whether there are heterogeneous effects of change feedback depending on its sign.

We can see that the overall positive effect of change feedback is driven by students who received negative change feedback. Columns 3 and 6 show that telling students who decreased their relative performance by how much their relative performance decreased increases their performance in the final test by 8.1 percentage points and by almost two thirds of a grade (0.6 grade points on a 6 point scale) as compared to their classmates who dropped in rank but received no feedback. These

effects are significant at the 1% level. Students who became worse and who received level feedback have a 5.3 percentage points and 0.4 grade points better outcome than students who received no feedback. These effects are significant at the 5% level.[17] Results when excluding class fixed effects, prior performance measures, and student characteristics can be found in Tables 3.12 and 3.13 in Appendix 3.7.5. Our results thus provide evidence in favor of hypothesis 1a: We find that, indeed, early level feedback significantly improves exam performance. However, our results contradict hypothesis 1b: Those who receive negative change feedback have a significant improvement in their performance (we expected it to worsen), while those who receive positive change feedback do not have a significant change in their performance, with a coefficient almost equal to zero (we expected it to improve). We will try to further explain the effects of change feedback in Section 3.5.4 by investigating whether it influenced student's effort effectiveness belief as described in Section 3.3.

**Change and level feedback given late** Table 3.4 presents the results with respect to the reference frame of feedback for classes that were treated immediately before the exam. As we saw above in Table 3.2 we did not find an overall significant effect of feedback given late. Looking at Table 3.4, neither feedback with a change frame nor feedback with a level frame had a significant effect on students knowledge at the exam as captured by their test scores. The first column shows that the overall effect of the change feedback is very close to zero. However, there seems to be heterogeneity in effects. The coefficient of the change feedback treatment dummy has a positive sign for students who improved (second column) and a negative sign for students who got worse (third column), although none of them are significant. However when looking at students grades we find a negative effect of 0.3 grade points of receiving negative change feedback immediately before the exam on one's grade. This effect is significant at the 5% level. This indicates that the effect of change feedback given immediately before the exam depends on whether the feedback is

[17] F-tests show that the coefficients of the change feedback and the level feedback in column 3 and column 6, respectively, are not significantly different from each other.

positive or negative and that while positive feedback tends to have no effect, negative feedback tends to have a negative effect.

Table 3.4: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Late Timing

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | -0.002 | 0.022 | -0.029 | 0.105 | -0.037 | 0.312** |
| | (0.018) | (0.040) | (0.023) | (0.106) | (0.240) | (0.149) |
| Level Frame | -0.022 | -0.009 | -0.023 | 0.176 | 0.025 | 0.271 |
| | (0.020) | (0.031) | (0.046) | (0.121) | (0.181) | (0.289) |
| Points Exam 1 | 0.125 | 0.105 | 0.382*** | -1.522*** | -2.171 | -2.832*** |
| | (0.122) | (0.293) | (0.129) | (0.580) | (1.399) | (0.813) |
| Points Exam 2 | 0.437*** | 0.429 | 0.256* | -2.818*** | -2.161 | -1.743** |
| | (0.110) | (0.269) | (0.137) | (0.499) | (1.353) | (0.797) |
| Female | -0.041 | -0.047 | -0.021 | 0.159 | 0.135 | 0.093 |
| | (0.031) | (0.039) | (0.028) | (0.160) | (0.162) | (0.205) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 159 | 76 | 83 | 159 | 76 | 83 |
| adj. $R^2$ | 0.361 | 0.204 | 0.456 | 0.393 | 0.289 | 0.436 |

*Note:* This table presents the effect of change frame and level frame feedback when given immediately before the exam using a linear regression model including class fixed effects.Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on class level and corrected using biased-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

Note that the signs of the coefficient for change feedback in the third and the sixth column of Table 3.4 is the reverse of the signs of the coefficient for change feedback in the respective columns of Table 3.3, indicating that while negative change feedback has a positive effect on educational outcomes when it is given 1-3 days before the exam, negative change feedback has the opposite effect when given immediately before the exam.

Additionally, all the signs of the coefficients of level feedback when it is given late

are the reverse to when it is given early: When level feedback is given 1-3 days before the exam we found it to have a generally positive effect while the same feedback when given immediately before the exam seems to have a negative effect, although it is smaller in magnitude and not significant. Results when excluding class fixed effects, prior performance measures, and student characteristics can be found in Tables 3.14 and 3.15 in Appendix 3.7.5. We do not find any significant effects of level feedback when given late feedback. However, the signs of the coefficients are in line with hypothesis 2a (when given late, level feedback has a negative effect). Although we do not find that student's exam scores are influenced by change feedback when it is given late, we find that feedback about negative changes negatively influences grades. This partly confirms hypothesis 2b (when given late, positive change feedback has a positive effect and negative change feedback has a negative effect).

### 3.5.4 Mechanisms

In this section we explore several behavioral mechanisms that might contribute to explaining our results. First, we look at whether the effects of feedback on outcomes can be explained by changes in the belief about the effectiveness of learning effort. Then, we will investigate whether feedback influenced emotions captured by students' state self-esteem.

**Effects of feedback on students' effort effectiveness belief** The "growth mindset" hypothesis described in section 3.3 predicts that making changes in past performance salient reinforces the belief that one's outcomes can be influenced by one's effort. We expected level feedback not to influence this belief. Table 3.16 in Appendix 3.7.6 shows that students who received change feedback report a weakly significant 0.16 standard deviations higher effort-effectiveness belief than the control group.[18] Furthermore, the results show that level feedback tends not to influence this belief.

[18]Note that, unlike in the regressions with test scores as dependent variables, we do not have pre-intervention information on students' effort effectiveness belief (or self-esteem) to control for level differences.

**Effects of feedback on students' self-esteem** We expected that level feedback as well as negative change feedback would on average be disappointing to students while positive change feedback would cheer them up. Table 3.17 in Appendix 3.7.6 shows that feedback tends to have a negative effect on students' self-esteem.

### 3.5.5 Sub-group Analyses

In the following we investigate whether effects of our feedback intervention is moderated by students' gender, preference for competition, math confidence and locus of control.

**Interaction with gender** Remarkably, as shown in Table 3.19 in Appendix 3.7.7, the overall positive effect of both change and level feedback in the early treatment is driven by the response of boys. Boys have 5.9 and 7.4 percentage points better results in the change and level treatments, respectively, than in the control group. At the same time, there is no significant difference for girls in any of the two treatment groups and the control group. This could possibly be driven by boys being more overconfident with respect to their prior knowledge than girls, as the literature suggests that (adult) males are more overconfident than (adult) females (Barber and Odean, 2001; Niederle and Vesterlund, 2007). Indeed, when we look at the effects of feedback on self-esteem of boys and girls separately, we find that it strongly reduces boys' self-esteem, while girls' self-esteem tends to be increased ( Table 3.18 in Appendix 3.7.6 ). This indicates that boys were on average disappointed by the feedback they received while girls were not.

Furthermore, looking at improvers and worseners separately reveals a positive effect of level feedback on boys who improved but no effect of any type of feedback on girls who improved, as F-tests show that the combined coefficients of the treatment dummies and the female indicators are not significantly different from zero. However, we find that both boys and girls respond positively to feedback about negative changes, as the coefficient of the interaction term of change feedback and female is

very small and insignificant. Analyses for classes that received feedback late reveal that neither the results of boys nor the results of girls are influenced by late feedback.

**Interactions with preference for competition and character traits** The psychological literature suggests that individual differences in character matter for how people react to (positive and negative) feedback (Ilgen et al., 1979). For example people with a more external locus of control may think that a bad outcome is due to factors they cannot control and may therefore not react to negative feedback by increasing their effort (Lam and Schaubroeck, 2000). People with high self-efficacy, i.e. a strong belief that they have the skills to complete a particular task, have been found to be more motivated by feedback than people who have low self-efficacy (Noe, 2000).

We do not find evidence that students' preference for competition (Table 3.21), confidence in math ability (Table 3.22) or locus of control (Table 3.23) explain their response to change or level feedback.

## 3.6 Conclusion

We investigated factors that may explain why feedback about past performance sometimes has positive and sometimes negative effects on performance. To do so we implemented a randomized feedback intervention in secondary schools. We varied the timing and reference frame of relative performance feedback to analyze its causal effect on performance in a high-stakes exam. With respect to timing, we compare students who received feedback either 1-3 days before the last math exam of the semester to students receiving the feedback immediately before the exam started. Concerning the reference frame of feedback students within the same class received either a level feedback, about their absolute rank in the preceding exam, or a change feedback, about their change in ranks between the two preceding exams, or no feedback

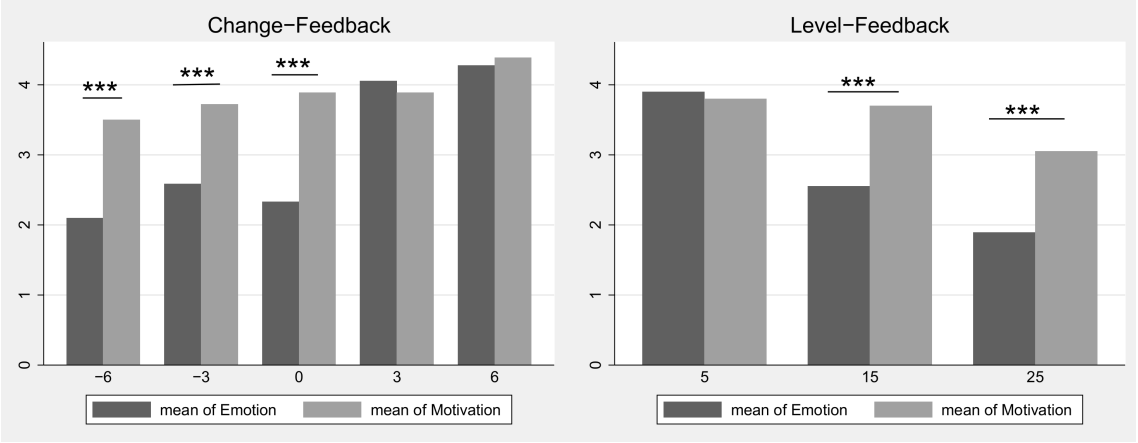
We expected that feedback affects both students expectations and emotions. We

find that level feedback and negative change feedback significantly improve outcomes in the final exam when given early but tend to decrease outcomes when given late. We do not find significant effects of positive change feedback. Our results also show that change (but not level) feedback strengthens the belief that one's outcomes can be influenced by one's effort and that feedback has an overall negative effect on students' emotions. Feedback has particularly strong effects on boys, while it is also boys' emotional state that is negatively affected by feedback. The results suggest that negatively surprising information about past performance may significantly improve performance in a high-stakes environment when it is given early enough, however, when it is given too late a negative emotional effect may dominate a positive incentive effect of information provision. Our results give interesting insights into the psychological and behavioral effects of relative performance feedback in an educational setting and has implications for the design of feedback in other situations where the ability to motivate people is crucial.

## 3.7 Appendix to Chapter 3

### 3.7.1 Results of Pre-experimental Survey

Figure 3.1: Pretest - Predicted Emotions and Motivation by Reference Frame of Feedback




*Note:* This graph shows the results of a pretest separately for change feedback (left) and level feedback (right). Dark bars are mean responses to the question *How do you think does Paul feel after reading the note?*, gray bars are mean responses to the question *How much do you think is Paul motivated to exert effort in the upcoming math exam?*. Both are measured on a 1 to 5 scale. Feedback notes in the pretest were varied such that students faced either a change in Paul's rank of -6, -3, 0, 3 or 6 or the ranks 5, 15 or 25. Differences between emotions and motivation were tested with a mean-comparison tests.

### 3.7.2 Feedback Notes

Figure 3.2: Feedback Note - Control Group [translated from German]

Dear [Student Name],

I wish you great success in your exam!

[Teacher Name]

Figure 3.3: Feedback Note - Change Frame Treatment [translated from German]

Dear [Student Name],

I compared the points of each student in the class in the last two exams.

**Relative to your classmates, you improved/worsened your performance in the last math exam by XX places.**

I wish you great success in your exam!

[Teacher Name]

Figure 3.4: Feedback Note - Level Frame Treatment [translated from German]

Dear [Student Name],

I looked at the points of each student in the class in the last exam.

**Relative to your classmates you achieved, with your performance in the last math exam, the XXth place.**

I wish you great success in your exam!

[Teacher Name]

### 3.7.3 Balance and Randomization Checks

Table 3.5: Treatment Observations

| | | **Class Level Randomization** | | |
|---|---|---|---|---|
| | | Late-Feedback Treatment | Early-Feedback Treatment | *Total Observations* |
| **Pupil Level Randomization** | Change Treatment | 57 | 59 | 116 |
| | Level Treatment | 61 | 64 | 125 |
| | Control Treatment | 56 | 55 | 111 |
| | *Total Observations* | 174 | 178 | *352* |

*Note:* This table summarizes the number of participants by treatment groups. In total, 352 children in 19 classes in 7 schools received parents' consent and participated.

Table 3.6: Randomization Check Class-Level Treatments

| | (1) Late-Feedback Treatment | (2) Early-Feedback Treatment | (3) Overall | (4) (1) vs. (2), p-value |
|---|---|---|---|---|
| Female Teacher | 0.793 | 0.781 | 0.787 | 0.781 |
| | (0.031) | (0.031) | (0.022) | |
| Class Size | 27.782 | 27.242 | 27.509 | 0.123 |
| | (0.244) | (0.250) | (0.175) | |
| Age | 23.667 | 24.708 | 24.193 | 0.363 |
| | (0.816) | (0.802) | (0.572) | |
| Points Exam1 | 0.712 | 0.681 | 0.696 | 0.105 |
| | (0.014) | (0.014) | (0.010) | |
| Points Exam2 | 0.719 | 0.730 | 0.725 | 0.554 |
| | (0.014) | (0.013) | (0.009) | |
| Rank Exam1 | 0.495 | 0.490 | 0.493 | 0.889 |
| | (0.022) | (0.021) | (0.015) | |
| Rank Exam2 | 0.467 | 0.493 | 0.481 | 0.399 |
| | (0.021) | (0.022) | (0.015) | |
| Change in Rank | 0.523 | −0.028 | 0.243 | 0.505 |
| | (0.592) | (0.577) | (0.413) | |
| Share Worsen | 0.506 | 0.455 | 0.480 | 0.343 |
| | (0.038) | (0.037) | (0.027) | |
| Share Participants | 0.775 | 0.703 | 0.739 | 0.000 |
| | (0.015) | (0.012) | (0.010) | |
| Female Pupil | 0.480 | 0.449 | 0.464 | 0.570 |
| | (0.038) | (0.037) | (0.027) | |
| Single Room | 0.655 | 0.596 | 0.625 | 0.370 |
| | (0.046) | (0.048) | (0.033) | |
| Internet | 1.115 | 1.022 | 1.068 | 0.366 |
| | (0.072) | (0.073) | (0.051) | |
| A-Level | 2.034 | 2.056 | 2.045 | 0.879 |
| | (0.103) | (0.099) | (0.071) | |
| Car | 1.333 | 1.303 | 1.318 | 0.785 |
| | (0.078) | (0.078) | (0.055) | |
| Siblings | 1.299 | 1.489 | 1.395 | 0.165 |
| | (0.094) | (0.099) | (0.068) | |
| Teacher Exp. | 9.902 | 12.833 | 11.513 | 0.008 |
| | (0.647) | (0.831) | (0.548) | |
| Books at Home | 1.983 | 2.140 | 2.063 | 0.314 |
| | (0.110) | (0.111) | (0.078) | |
| $N$ | 174 | 178 | 352 | |
| Proportion | 0.494 | 0.506 | 1.000 | |

*Note:* Standard errors in parentheses.

Table 3.7: Randomization Check Student-Level Treatments - Early Timing

| | (1) Control | (2) Change | (3) Level | (4) Overall | (5) (1) vs. (2), p-value | (6) (1) vs. (3), p-value | (7) (2) vs. (3), p-value |
|---|---|---|---|---|---|---|---|
| Female Teacher | 0.782 | 0.780 | 0.781 | 0.781 | 0.978 | 0.994 | 0.983 |
| | (0.056) | (0.054) | (0.052) | (0.031) | | | |
| Class Size | 27.255 | 27.322 | 27.156 | 27.242 | 0.914 | 0.874 | 0.784 |
| | (0.452) | (0.429) | (0.424) | (0.250) | | | |
| Age | 23.750 | 23.415 | 23.286 | 23.478 | 0.820 | 0.761 | 0.930 |
| | (1.069) | (1.005) | (1.080) | (0.604) | | | |
| Points Exam1 | 0.691 | 0.661 | 0.690 | 0.681 | 0.422 | 0.967 | 0.407 |
| | (0.025) | (0.027) | (0.021) | (0.014) | | | |
| Points Exam2 | 0.733 | 0.732 | 0.727 | 0.730 | 0.976 | 0.845 | 0.866 |
| | (0.023) | (0.022) | (0.021) | (0.013) | | | |
| Rank Exam1 | 0.472 | 0.510 | 0.488 | 0.490 | 0.487 | 0.751 | 0.678 |
| | (0.038) | (0.038) | (0.035) | (0.021) | | | |
| Rank Exam2 | 0.482 | 0.487 | 0.508 | 0.493 | 0.934 | 0.637 | 0.687 |
| | (0.041) | (0.038) | (0.037) | (0.022) | | | |
| Change in Rank | −0.382 | 0.763 | −0.453 | −0.028 | 0.424 | 0.959 | 0.400 |
| | (0.959) | (1.048) | (0.988) | (0.577) | | | |
| Share Worsen | 0.455 | 0.458 | 0.453 | 0.455 | 0.974 | 0.988 | 0.960 |
| | (0.068) | (0.065) | (0.063) | (0.037) | | | |
| Share Participants | 0.710 | 0.701 | 0.699 | 0.703 | 0.751 | 0.706 | 0.959 |
| | (0.022) | (0.021) | (0.020) | (0.012) | | | |
| Female Pupil | 0.418 | 0.424 | 0.500 | 0.449 | 0.953 | 0.376 | 0.401 |
| | (0.067) | (0.065) | (0.063) | (0.037) | | | |
| Single Room | 0.765 | 0.759 | 0.707 | 0.742 | 0.948 | 0.500 | 0.536 |
| | (0.060) | (0.059) | (0.060) | (0.034) | | | |
| Internet | 1.100 | 1.315 | 1.241 | 1.222 | 0.129 | 0.345 | 0.628 |
| | (0.096) | (0.102) | (0.111) | (0.060) | | | |
| A-level | 2.347 | 2.453 | 2.582 | 2.465 | 0.500 | 0.130 | 0.292 |
| | (0.132) | (0.088) | (0.085) | (0.059) | | | |
| Car | 1.471 | 1.648 | 1.518 | 1.547 | 0.255 | 0.731 | 0.363 |
| | (0.106) | (0.113) | (0.088) | (0.059) | | | |
| Siblings | 1.462 | 1.288 | 1.466 | 1.407 | 0.170 | 0.975 | 0.145 |
| | (0.093) | (0.084) | (0.086) | (0.051) | | | |
| Teacher Exp. | 12.638 | 12.980 | 12.870 | 12.833 | 0.870 | 0.910 | 0.957 |
| | (1.471) | (1.466) | (1.409) | (0.831) | | | |
| Books at Home | 2.231 | 2.679 | 2.379 | 2.429 | 0.057 | 0.481 | 0.205 |
| | (0.144) | (0.182) | (0.151) | (0.093) | | | |
| $N$ | 55 | 59 | 64 | 178 | | | |
| Proportion | 0.309 | 0.331 | 0.360 | 1.000 | | | |

*Note:* Standard errors in parentheses.

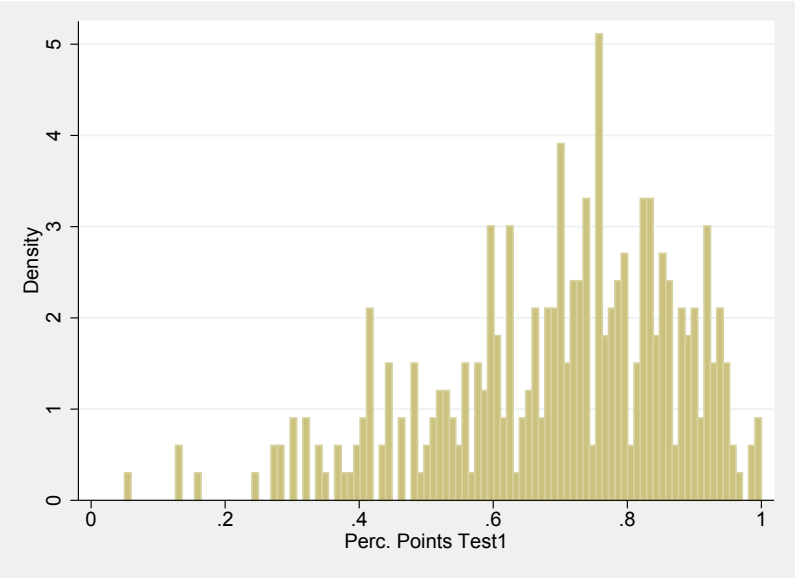
Table 3.8: Randomization Check Student-Level Treatments - Late Timing

| | (1) Control | (2) Change | (3) Level | (4) Overall | (5) (1) vs. (2), p-value | (6) (1) vs. (3), p-value | (7) (2) vs. (3), p-value |
|---|---|---|---|---|---|---|---|
| Female Teacher | 0.782 | 0.789 | 0.800 | 0.791 | 0.922 | 0.813 | 0.889 |
| | (0.056) | (0.054) | (0.052) | (0.031) | | | |
| Class Size | 27.782 | 27.877 | 27.667 | 27.773 | 0.874 | 0.852 | 0.730 |
| | (0.429) | (0.421) | (0.437) | (0.247) | | | |
| Age | 22.667 | 22.075 | 22.429 | 22.382 | 0.712 | 0.885 | 0.823 |
| | (1.174) | (1.086) | (1.136) | (0.650) | | | |
| Points Exam1 | 0.745 | 0.708 | 0.703 | 0.718 | 0.264 | 0.179 | 0.871 |
| | (0.022) | (0.024) | (0.022) | (0.013) | | | |
| Points Exam2 | 0.730 | 0.712 | 0.717 | 0.719 | 0.581 | 0.681 | 0.881 |
| | (0.024) | (0.024) | (0.023) | (0.014) | | | |
| Rank Exam1 | 0.438 | 0.502 | 0.522 | 0.489 | 0.253 | 0.105 | 0.706 |
| | (0.039) | (0.040) | (0.034) | (0.022) | | | |
| Rank Exam2 | 0.457 | 0.470 | 0.475 | 0.467 | 0.800 | 0.728 | 0.924 |
| | (0.038) | (0.036) | (0.036) | (0.021) | | | |
| Change in Rank | −0.600 | 0.842 | 1.250 | 0.523 | 0.342 | 0.190 | 0.777 |
| | (1.044) | (1.091) | (0.943) | (0.592) | | | |
| Share Worsen | 0.527 | 0.544 | 0.467 | 0.512 | 0.862 | 0.520 | 0.408 |
| | (0.068) | (0.067) | (0.065) | (0.038) | | | |
| Share Participants | 0.778 | 0.772 | 0.770 | 0.773 | 0.861 | 0.812 | 0.953 |
| | (0.026) | (0.026) | (0.025) | (0.015) | | | |
| Female Pupil | 0.418 | 0.544 | 0.475 | 0.480 | 0.186 | 0.549 | 0.460 |
| | (0.067) | (0.067) | (0.066) | (0.038) | | | |
| Single Room | 0.745 | 0.811 | 0.804 | 0.787 | 0.421 | 0.474 | 0.919 |
| | (0.062) | (0.054) | (0.054) | (0.032) | | | |
| Internet | 1.235 | 1.255 | 1.411 | 1.304 | 0.898 | 0.220 | 0.278 |
| | (0.107) | (0.108) | (0.095) | (0.059) | | | |
| A-level | 2.511 | 2.320 | 2.604 | 2.480 | 0.251 | 0.518 | 0.059 |
| | (0.113) | (0.119) | (0.091) | (0.063) | | | |
| Car | 1.431 | 1.491 | 1.655 | 1.528 | 0.694 | 0.168 | 0.309 |
| | (0.106) | (0.106) | (0.120) | (0.064) | | | |
| Siblings | 1.220 | 1.245 | 1.268 | 1.245 | 0.866 | 0.742 | 0.874 |
| | (0.108) | (0.104) | (0.097) | (0.059) | | | |
| Teacher Exp. | 9.795 | 9.725 | 9.930 | 9.820 | 0.966 | 0.933 | 0.897 |
| | (1.159) | (1.132) | (1.098) | (0.647) | | | |
| Books at Home | 2.160 | 2.189 | 2.382 | 2.247 | 0.900 | 0.361 | 0.409 |
| | (0.167) | (0.155) | (0.173) | (0.095) | | | |
| $N$ | 55 | 57 | 60 | 172 | | | |
| Proportion | 0.320 | 0.331 | 0.349 | 1.000 | | | |

*Note:* Standard errors in parentheses.

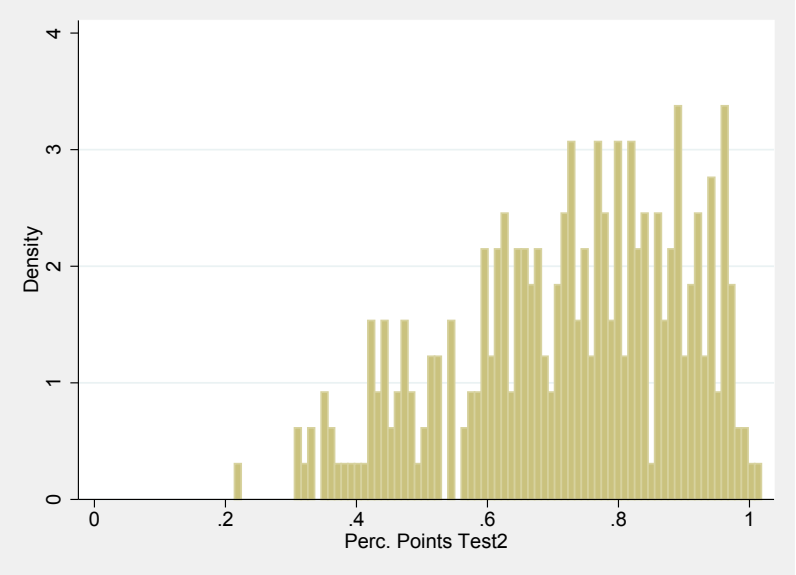
### 3.7.4 Graphs

Figure 3.5: Distribution of points in Test 1



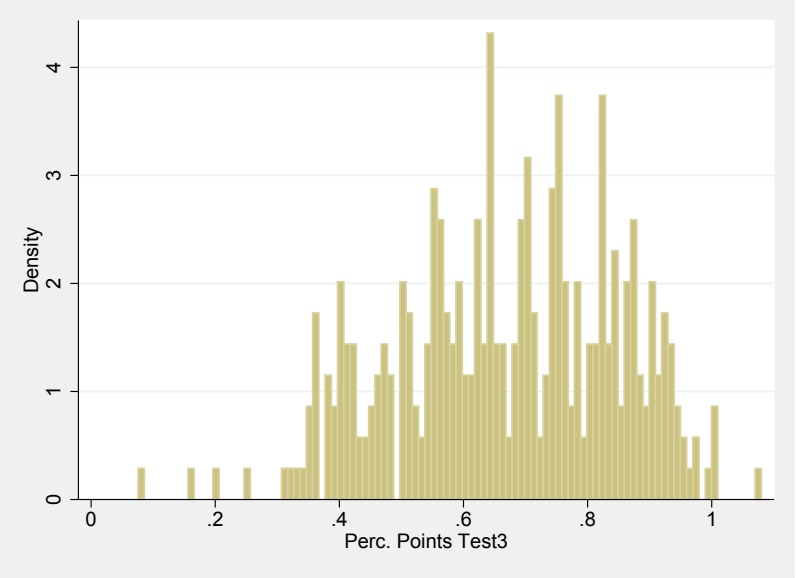

*Note:* This graph shows the distribution of points in test 1.

Figure 3.6: Distribution of points in Test 2



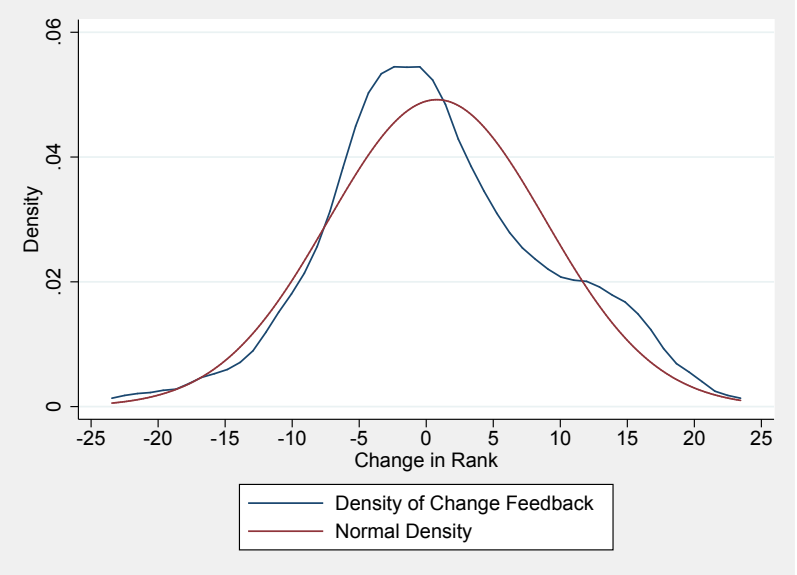

*Note:* This graph shows the distribution of points in test 2.

Figure 3.7: Distribution of points in Test 3



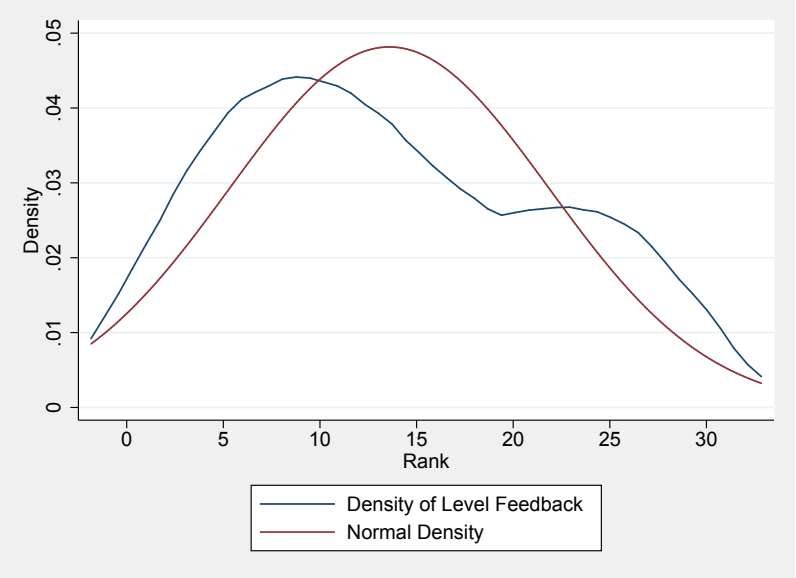

*Note:* This graph shows the distribution of points in test 3.

Figure 3.8: Feedback in Change Frame Treatment




*Note:* This graph shows kernel density estimates for the feedback students received in the Change Treatment.

Figure 3.9: Feedback in Level Frame Treatment




*Note:* This graph shows kernel density estimates for the feedback students received in the Level Treatment.

### 3.7.5 Check for Spillovers and Robustness Checks

Table 3.9: Check for Spillover Effects

| | (1) Points in Exam 3 (Control Group) | (2) Grade in Exam 3 (Control Group) |
|---|---|---|
| Early Timing | 0.042 | -0.161 |
| | (0.033) | (0.193) |
| Points Exam 1 | 0.271 | -2.452*** |
| | (0.164) | (0.847) |
| Points Exam 2 | 0.395** | -2.512*** |
| | (0.159) | (0.860) |
| Female | 0.014 | -0.116 |
| | (0.027) | (0.184) |
| SchoolFE | Yes | Yes |
| Pupil Controls | Yes | Yes |
| $N$ | 101 | 101 |
| adj. $R^2$ | 0.274 | 0.370 |

*Note:* This table presents the differences in outcomes of the control groups of classes who had the intervention early and classes who had the intervention late. In column 1 the dependent variable percentage points in exam 3, while in column 2 the dependent variable is grades in exam 3. Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters 19. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.10: Robustness Checks - Class-Level Treatments - Points

| | Dep. Var.: Points in Exam 3 | | | | | |
|---|---|---|---|---|---|---|
| | If Early Timing | If Early Timing | If Early Timing | If Late Timing | If Late Timing | If Late Timing |
| Feedback | 0.026* | 0.026 | 0.032* | -0.032 | -0.029* | -0.013 |
| | (0.015) | (0.016) | (0.016) | (0.019) | (0.017) | (0.014) |
| Points Exam 1 | | | 0.314*** | | | 0.182 |
| | | | (0.044) | | | (0.113) |
| Points Exam 2 | | | 0.281*** | | | 0.460*** |
| | | | (0.107) | | | (0.106) |
| Female | | | -0.007 | | | -0.042 |
| | | | (0.030) | | | (0.027) |
| ClassFE | No | Yes | No | No | Yes | No |
| Pupil Controls | No | No | Yes | No | No | Yes |
| $N$ | 160 | 160 | 160 | 159 | 159 | 159 |
| adj. $R^2$ | -0.001 | 0.160 | 0.411 | 0.000 | 0.141 | 0.315 |

*Note:* This table presents the effect of feedback timing on performance in the last exam using a linear regression model. Dependent variable: percentage points in exam 3. Models 1 and 4 do not contain any control variables. Models 2 and 5 contain class fixed effects but no other control variables. Models 3 and 6 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain class fixed effects. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 10 in models 1, 2, and 3 and 9 in models 4, 5, and 6. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.11: Robustness Checks - Class-Level Treatments - Grade

| | Dep. Var.: Grade in Exam 3 | | | | | |
|---|---|---|---|---|---|---|
| | If Early Timing | If Early Timing | If Early Timing | If Late Timing | If Late Timing | If Late Timing |
| Feedback | -0.133 | -0.144 | -0.187** | 0.282* | 0.253* | 0.175** |
| | (0.105) | (0.112) | (0.091) | (0.146) | (0.129) | (0.083) |
| Points Exam 1 | | | -2.218*** | | | -1.870*** |
| | | | (0.441) | | | (0.536) |
| Points Exam 2 | | | -1.855** | | | -2.645*** |
| | | | (0.838) | | | (0.459) |
| Female | | | 0.062 | | | 0.153 |
| | | | (0.198) | | | (0.151) |
| ClassFE | No | Yes | No | No | Yes | No |
| Pupil Controls | No | No | Yes | No | No | Yes |
| $N$ | 160 | 160 | 160 | 159 | 159 | 159 |
| adj. $R^2$ | -0.003 | 0.158 | 0.403 | 0.006 | 0.117 | 0.358 |

*Note:* This table presents the effect of feedback timing on performance in the last exam using a linear regression model. Dependent variable: grade in exam 3. Models 1 and 4 do not contain any control variables. Models 2 and 5 contain class fixed effects but no other control variables. Models 3 and 6 control for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but do not contain class fixed effects. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 10 in models 1, 2, and 3 and 9 in models 4, 5, and 6. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.12: Robustness Checks - Student-Level Treatments in Early Timing - Points

| | Dep. Var.: Points in Exam 3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | If Improved | If Improved | If Improved | If Worsened | If Worsened | If Worsened |
| Change Frame | 0.027 | 0.024 | 0.033 | -0.031 | -0.033 | -0.014 | 0.095*** | 0.094*** | 0.089** |
| | (0.019) | (0.020) | (0.020) | (0.053) | (0.055) | (0.048) | (0.024) | (0.018) | (0.035) |
| Level Frame | 0.025 | 0.029 | 0.030* | -0.009 | 0.001 | 0.016 | 0.066*** | 0.073** | 0.053 |
| | (0.019) | (0.020) | (0.017) | (0.036) | (0.041) | (0.039) | (0.018) | (0.029) | (0.034) |
| Points Exam 1 | | | 0.314*** | | | 0.191 | | | 0.511*** |
| | | | (0.044) | | | (0.132) | | | (0.150) |
| Points Exam 2 | | | 0.280*** | | | 0.480*** | | | 0.081 |
| | | | (0.106) | | | (0.084) | | | (0.187) |
| Female | | | -0.006 | | | -0.005 | | | -0.009 |
| | | | (0.030) | | | (0.046) | | | (0.038) |
| ClassFE | No | Yes | No | No | Yes | No | No | Yes | No |
| Pupil Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| $N$ | 160 | 160 | 160 | 87 | 87 | 87 | 73 | 73 | 73 |
| adj. $R^2$ | -0.008 | 0.155 | 0.407 | -0.019 | 0.082 | 0.363 | 0.028 | 0.264 | 0.482 |

*Note:* This table presents the effect of change frame and level frame feedback when given 1-3 days in advance using a linear regression model. Dependent variable: percentage points in exam 3. Models 1, 4, and 7 does not contain any control variables. Models 2, 5, and 8 contains class fixed effects but no other control variables. Models 3, 6, and 9 controls for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but does not contain class fixed effects. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.13: Robustness Checks - Student-Level Treatments in Early Timing - Grade

| | Dep. Var.: Grade in Exam 3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | If Improved | If Improved | If Improved | If Worsened | If Worsened | If Worsened |
| Change Frame | -0.124 | -0.113 | -0.177 | 0.310 | 0.310 | 0.173 | -0.634*** | -0.663*** | -0.599** |
| | (0.136) | (0.131) | (0.120) | (0.378) | (0.377) | (0.337) | (0.213) | (0.124) | (0.268) |
| Level Frame | -0.142 | -0.172 | -0.196* | 0.074 | -0.000 | -0.100 | -0.405*** | -0.499*** | -0.343 |
| | (0.126) | (0.134) | (0.101) | (0.250) | (0.280) | (0.253) | (0.152) | (0.172) | (0.271) |
| Points Exam 1 | | | -2.214*** | | | -1.524 | | | -3.652*** |
| | | | (0.439) | | | (0.991) | | | (1.096) |
| Points Exam 2 | | | -1.859** | | | -3.069*** | | | -0.342 |
| | | | (0.832) | | | (0.796) | | | (1.512) |
| Female | | | 0.063 | | | 0.023 | | | 0.099 |
| | | | (0.201) | | | (0.289) | | | (0.262) |
| ClassFE | No | Yes | No | No | Yes | No | No | Yes | No |
| Pupil Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| $N$ | 160 | 160 | 160 | 87 | 87 | 87 | 73 | 73 | 73 |
| adj. $R^2$ | -0.009 | 0.153 | 0.399 | -0.012 | 0.091 | 0.382 | 0.025 | 0.284 | 0.440 |

*Note:* This table presents the effect of change frame and level frame feedback when given 1-3 days in advance using a linear regression model. Dependent variable:grade in exam 3. Models 1, 4, and 7 does not contain any control variables. Models 2, 5, and 8 contains class fixed effects but no other control variables. Models 3, 6, and 9 controls for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but does not contain class fixed effects. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.14: Robustness Checks - Student-Level Treatments in Late Timing - Points

| | Dep. Var.: Points in Exam 3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | If Improved | If Improved | If Improved | If Worsened | If Worsened | If Worsened |
| Change Frame | -0.018 | -0.018 | -0.001 | -0.028 | -0.012 | 0.015 | -0.006 | -0.025 | -0.011 |
| | (0.023) | (0.021) | (0.017) | (0.030) | (0.031) | (0.037) | (0.044) | (0.030) | (0.025) |
| Level Frame | -0.044** | -0.040** | -0.025 | -0.035 | -0.016 | -0.024 | -0.059 | -0.064* | -0.015 |
| | (0.022) | (0.019) | (0.019) | (0.028) | (0.028) | (0.034) | (0.043) | (0.036) | (0.038) |
| Points Exam 1 | | | 0.180 | | | 0.139 | | | 0.391*** |
| | | | (0.114) | | | (0.258) | | | (0.126) |
| Points Exam 2 | | | 0.462*** | | | 0.435* | | | 0.346*** |
| | | | (0.107) | | | (0.248) | | | (0.116) |
| Female | | | -0.044 | | | -0.055 | | | -0.027 |
| | | | (0.029) | | | (0.043) | | | (0.029) |
| ClassFE | No | Yes | No | No | Yes | No | No | Yes | No |
| Pupil Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| $N$ | 159 | 159 | 159 | 76 | 76 | 76 | 83 | 83 | 83 |
| adj. $R^2$ | -0.001 | 0.138 | 0.314 | -0.019 | 0.043 | 0.183 | -0.000 | 0.242 | 0.354 |

*Note:* This table presents the effect of change frame and level frame feedback when given 1-3 days in advance using a linear regression model. Dependent variable: percentage points in exam 3. Models 1, 4, and 7 does not contain any control variables. Models 2, 5, and 8 contains class fixed effects but no other control variables. Models 3, 6, and 9 controls for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but does not contain class fixed effects. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.15: Robustness Checks - Student-Level Treatments in Late Timing - Grade

| | Dep. Var.: Grade in Exam 3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | If Improved | If Improved | If Improved | If Worsened | If Worsened | If Worsened |
| Change Frame | 0.213 | 0.204 | 0.124 | 0.247 | 0.182 | -0.022 | 0.166 | 0.234 | 0.266 |
| | (0.177) | (0.156) | (0.105) | (0.222) | (0.237) | (0.209) | (0.339) | (0.240) | (0.177) |
| Level Frame | 0.347** | 0.299** | 0.223** | 0.116 | 0.019 | 0.092 | 0.614* | 0.545* | 0.324 |
| | (0.162) | (0.144) | (0.104) | (0.170) | (0.159) | (0.177) | (0.335) | (0.284) | (0.251) |
| Points Exam 1 | | | -1.860*** | | | -2.196* | | | -2.888*** |
| | | | (0.543) | | | (1.225) | | | (0.853) |
| Points Exam 2 | | | -2.652*** | | | -1.923 | | | -1.956*** |
| | | | (0.462) | | | (1.164) | | | (0.630) |
| Female | | | 0.158 | | | 0.165 | | | 0.098 |
| | | | (0.156) | | | (0.171) | | | (0.214) |
| ClassFE | No | Yes | No | No | Yes | No | No | Yes | No |
| Pupil Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| $N$ | 159 | 159 | 159 | 76 | 76 | 76 | 83 | 83 | 83 |
| adj. $R^2$ | 0.002 | 0.113 | 0.355 | -0.019 | -0.012 | 0.282 | 0.023 | 0.200 | 0.375 |

*Note:* This table presents the effect of change frame and level frame feedback when given 1-3 days in advance using a linear regression model. Dependent variable:grade in exam 3. Models 1, 4, and 7 does not contain any control variables. Models 2, 5, and 8 contains class fixed effects but no other control variables. Models 3, 6, and 9 controls for percentage points exam 1, percentage points exam 2, gender, own room, foreign name, and siblings but does not contain class fixed effects. Standard errors are reported in parentheses, clustered on class level and corrected using bias-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

### 3.7.6 Mechanisms: Effort-effectiveness Belief and Self-esteem

Table 3.16: Change Frame vs. Level Frame vs. Control - Dep. var. effort effectiveness belief

| | (1) All | (2) If Improved | (3) If Worsened |
|---|---|---|---|
| Change Frame | 0.168* | 0.274 | 0.228 |
| | (0.093) | (0.216) | (0.182) |
| Level Frame | 0.017 | 0.144 | -0.065 |
| | (0.155) | (0.216) | (0.241) |
| Points Exam 1 | 1.003*** | 1.693*** | 0.922 |
| | (0.272) | (0.618) | (1.621) |
| Points Exam 2 | 1.273** | 0.187 | 1.259 |
| | (0.559) | (1.095) | (1.309) |
| Female | -0.079 | -0.147 | 0.063 |
| | (0.118) | (0.102) | (0.262) |
| ClassFE | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes |
| $N$ | 161 | 88 | 73 |
| adj. $R^2$ | 0.0868 | 0.1192 | -0.0763 |

*Note:* This table presents the effect of change frame and level frame feedback on effectiveness belief using a linear regression model including class fixed effects. Model 1 present results for the whole sample in each early and late treatment classes, model 2 present results for students who improved, and model 3 present results for students who worsened their performance from the second last to the last exam. Dependent variable: effort-effectiveness belief (standardized to zero mean and unit standard deviation). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.17: Change Frame vs. Level Frame vs. Control - Dep. var. state self-esteem

| | (1) All | (2) If Improved | (3) If Worsened |
|---|---|---|---|
| Change Frame | -0.206 | -0.437* | -0.075 |
| | (0.130) | (0.232) | (0.249) |
| Level Frame | -0.280* | -0.442* | -0.058 |
| | (0.142) | (0.232) | (0.263) |
| Points Exam 1 | 0.715 | 1.593 | 0.189 |
| | (0.673) | (0.991) | (2.276) |
| Points Exam 2 | 1.507*** | 0.251 | 1.640 |
| | (0.535) | (1.234) | (1.699) |
| Female | -0.113 | -0.011 | -0.028 |
| | (0.168) | (0.327) | (0.201) |
| ClassFE | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes |
| $N$ | 151 | 81 | 70 |
| adj. $R^2$ | 0.1321 | 0.1478 | 0.0568 |

*Note:* This table presents the effect of change frame and level frame feedback on state self-esteem using a linear regression model including class fixed effects. Model 1 present results for the whole sample in each early and late treatment classes, model 2 present results for students who improved, and model 3 present results for students who worsened their performance from the second last to the last exam. Dependent variable: state-self esteem (standardized to zero mean and unit standard deviation). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.18: Change Frame vs. Level Frame vs. Control - Dep. var. state self-esteem (by gender)

| | (1) Boys | (2) Girls |
|---|---|---|
| Change Frame | -0.549** | 0.439* |
| | (0.270) | (0.250) |
| Level Frame | -0.464*** | -0.095 |
| | (0.150) | (0.258) |
| Points Exam 1 | 0.867 | 0.771 |
| | (1.470) | (0.740) |
| Points Exam 2 | 2.041* | 0.712 |
| | (1.143) | (0.493) |
| ClassFE | Yes | Yes |
| Pupil Controls | Yes | Yes |
| $N$ | 80 | 71 |
| adj. $R^2$ | 0.1819 | 0.1193 |

*Note:* This table presents the effect of change frame and level frame feedback on state self-esteem using a linear regression model including class fixed effects. Model 1 present results for boys, model 2 present results for girls. Dependent variable: state self-esteem (standardized to zero mean and unit standard deviation). Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

### 3.7.7 Sub-group Analyses

Table 3.19: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with gender)

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | 0.059** | 0.050 | 0.084** | -0.372** | -0.185 | -0.688** |
| | (0.024) | (0.054) | (0.041) | (0.149) | (0.380) | (0.318) |
| Change Frame X Female | -0.051* | -0.119* | -0.006 | 0.351* | 0.592* | 0.246 |
| | (0.028) | (0.064) | (0.058) | (0.199) | (0.350) | (0.420) |
| Level Frame | 0.074*** | 0.096*** | 0.066 | -0.451*** | -0.609*** | -0.437 |
| | (0.011) | (0.031) | (0.049) | (0.079) | (0.216) | (0.314) |
| Level Frame X Female | -0.073** | -0.161*** | -0.023 | 0.417** | 0.973*** | 0.122 |
| | (0.031) | (0.053) | (0.074) | (0.202) | (0.286) | (0.481) |
| Points Exam 1 | 0.363*** | 0.326** | 0.483*** | -2.605*** | -2.430** | -3.701*** |
| | (0.044) | (0.123) | (0.161) | (0.425) | (0.939) | (1.158) |
| Points Exam 2 | 0.293*** | 0.337*** | 0.148 | -1.988*** | -2.297*** | -0.515 |
| | (0.074) | (0.117) | (0.147) | (0.541) | (0.842) | (1.129) |
| Female | 0.049 | 0.094* | 0.030 | -0.289 | -0.569* | -0.199 |
| | (0.038) | (0.054) | (0.055) | (0.224) | (0.311) | (0.363) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 160 | 87 | 73 | 160 | 87 | 73 |
| adj. $R^2$ | 0.518 | 0.443 | 0.597 | 0.543 | 0.494 | 0.620 |

*Note:* This table presents the effect of change frame and level frame feedback interacted with students' gender when given 1-3 days in advance using a linear regression model including class fixed effects.Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings, grade in exam 1 (5 categories), grade in exam 2 (5 categories). Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.20: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Late Timing (Interaction with gender)

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | -0.013 | -0.004 | -0.030 | 0.210 | 0.053 | 0.316 |
| | (0.030) | (0.063) | (0.037) | (0.166) | (0.413) | (0.196) |
| Change Frame X Female | 0.020 | 0.072 | 0.006 | -0.208 | -0.363 | -0.036 |
| | (0.062) | (0.070) | (0.068) | (0.425) | (0.466) | (0.466) |
| Level Frame | -0.014 | -0.041 | 0.018 | 0.186 | 0.355 | 0.020 |
| | (0.024) | (0.032) | (0.064) | (0.191) | (0.245) | (0.386) |
| Level Frame X Female | -0.017 | 0.086 | -0.073 | -0.024 | -0.840 | 0.443 |
| | (0.050) | (0.075) | (0.092) | (0.382) | (0.523) | (0.640) |
| Points Exam 1 | 0.122 | 0.094 | 0.412*** | -1.522*** | -2.265 | -3.014*** |
| | (0.130) | (0.321) | (0.129) | (0.578) | (1.487) | (0.779) |
| Points Exam 2 | 0.435*** | 0.433 | 0.227 | -2.790*** | -2.057 | -1.571** |
| | (0.114) | (0.286) | (0.141) | (0.491) | (1.408) | (0.783) |
| Female | -0.041 | -0.104** | 0.001 | 0.234 | 0.588** | -0.041 |
| | (0.030) | (0.043) | (0.060) | (0.251) | (0.267) | (0.428) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 159 | 76 | 83 | 159 | 76 | 83 |
| adj. $R^2$ | 0.353 | 0.187 | 0.451 | 0.386 | 0.285 | 0.427 |

*Note:* This table presents the effect of change frame and level frame feedback interacted with students' gender when given immediately before the exam using a linear regression model including class fixed effects.Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 9. * p<0.10, ** p<0.05, *** p<0.01.

Table 3.21: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with preference for competition)

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | 0.038 | -0.073 | 0.127** | -0.318 | 0.381 | -0.949*** |
| | (0.041) | (0.067) | (0.054) | (0.265) | (0.412) | (0.292) |
| Change Frame X High Comp. | -0.005 | 0.100 | -0.096 | 0.193 | -0.436 | 0.766** |
| | (0.062) | (0.063) | (0.070) | (0.352) | (0.388) | (0.367) |
| Level Frame | 0.047 | -0.017 | 0.122 | -0.301 | 0.053 | -0.818* |
| | (0.054) | (0.074) | (0.077) | (0.343) | (0.446) | (0.459) |
| Level Frame X High Comp. | -0.020 | 0.039 | -0.119 | 0.110 | -0.216 | 0.741 |
| | (0.071) | (0.075) | (0.098) | (0.470) | (0.458) | (0.603) |
| High Competitiveness | -0.031 | -0.093* | 0.045 | 0.157 | 0.509* | -0.306 |
| | (0.043) | (0.049) | (0.056) | (0.279) | (0.290) | (0.340) |
| Points Exam 1 | 0.334*** | 0.288* | 0.537*** | -2.401*** | -2.160* | -4.075*** |
| | (0.059) | (0.159) | (0.110) | (0.527) | (1.141) | (0.768) |
| Points Exam 2 | 0.317*** | 0.373*** | 0.111 | -2.177*** | -2.604*** | -0.245 |
| | (0.065) | (0.132) | (0.113) | (0.497) | (0.975) | (0.826) |
| Female | -0.002 | -0.016 | 0.015 | 0.028 | 0.057 | -0.040 |
| | (0.026) | (0.055) | (0.021) | (0.165) | (0.340) | (0.113) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Control | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 160 | 87 | 73 | 160 | 87 | 73 |
| adj. $R^2$ | 0.519 | 0.424 | 0.626 | 0.546 | 0.474 | 0.650 |

*Note:* This table presents the effect of change frame and level frame feedback interacted with students' preference for competition when given 1-3 days in advance using a linear regression model including class fixed effects.Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, ***p<0.01.

Table 3.22: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with confidence in math ability)

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | 0.052 | 0.004 | 0.084** | -0.454* | -0.005 | -0.836*** |
| | (0.038) | (0.078) | (0.039) | (0.245) | (0.489) | (0.222) |
| Change Frame X High Math Conf. | -0.025 | 0.006 | -0.005 | 0.407 | 0.038 | 0.436 |
| | (0.062) | (0.096) | (0.074) | (0.349) | (0.574) | (0.449) |
| Level Frame | 0.062* | 0.064 | 0.044* | -0.486* | -0.445 | -0.418** |
| | (0.037) | (0.097) | (0.025) | (0.258) | (0.616) | (0.178) |
| Level Frame X High Math Conf. | -0.038 | -0.048 | 0.022 | 0.375 | 0.333 | 0.058 |
| | (0.050) | (0.099) | (0.067) | (0.316) | (0.618) | (0.423) |
| High Math Confidence | 0.009 | -0.026 | 0.005 | -0.150 | 0.151 | -0.128 |
| | (0.036) | (0.071) | (0.071) | (0.235) | (0.442) | (0.470) |
| Points Exam 1 | 0.363*** | 0.330** | 0.468*** | -2.669*** | -2.463** | -3.710*** |
| | (0.047) | (0.136) | (0.125) | (0.450) | (1.021) | (0.959) |
| Points Exam 2 | 0.311*** | 0.398*** | 0.162 | -2.115*** | -2.714** | -0.753 |
| | (0.063) | (0.149) | (0.147) | (0.476) | (1.093) | (1.087) |
| Female | 0.006 | -0.007 | 0.021 | -0.023 | 0.004 | -0.103 |
| | (0.027) | (0.053) | (0.023) | (0.173) | (0.338) | (0.150) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Control | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 160 | 87 | 73 | 160 | 87 | 73 |
| adj. $R^2$ | 0.509 | 0.413 | 0.591 | 0.542 | 0.471 | 0.619 |

*Note:* This table presents the effect of change frame and level frame feedback interacted with students' confidence in math ability when given 1-3 days in advance using a linear regression model including class fixed effects.Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, ***

Table 3.23: Change Frame vs. Level Frame vs. Control - Class-Level Treatment: Early Timing (Interaction with locus of control)

| | Dep. Var.: Points in Exam 3 | | | Dep. Var.: Grade in Exam 3 | | |
|---|---|---|---|---|---|---|
| | All | If Improved | If Worsened | All | If Improved | If Worsened |
| Change Frame | 0.057 | 0.024 | 0.084 | -0.374 | -0.134 | -0.600 |
| | (0.051) | (0.102) | (0.058) | (0.354) | (0.730) | (0.367) |
| Change Frame X Internal LOC | -0.034 | -0.032 | -0.003 | 0.271 | 0.269 | 0.023 |
| | (0.076) | (0.113) | (0.074) | (0.481) | (0.762) | (0.454) |
| Level Frame | 0.032 | 0.045 | 0.019 | -0.262 | -0.377 | -0.151 |
| | (0.048) | (0.083) | (0.065) | (0.308) | (0.581) | (0.420) |
| Level X Internal LOC | 0.019 | -0.027 | 0.064 | -0.030 | 0.298 | -0.433 |
| | (0.065) | (0.094) | (0.100) | (0.391) | (0.659) | (0.586) |
| Internal LOC | 0.025 | 0.047 | -0.005 | -0.200 | -0.419 | 0.111 |
| | (0.053) | (0.096) | (0.067) | (0.344) | (0.652) | (0.416) |
| Points Exam 1 | 0.344*** | 0.327* | 0.450*** | -2.515*** | -2.446** | -3.585*** |
| | (0.055) | (0.167) | (0.134) | (0.485) | (1.208) | (1.019) |
| Points Exam 2 | 0.293*** | 0.335** | 0.175 | -1.966*** | -2.306** | -0.640 |
| | (0.063) | (0.145) | (0.130) | (0.495) | (1.070) | (1.011) |
| Female | 0.002 | -0.008 | 0.015 | 0.007 | 0.009 | -0.043 |
| | (0.031) | (0.052) | (0.019) | (0.203) | (0.324) | (0.098) |
| ClassFE | Yes | Yes | Yes | Yes | Yes | Yes |
| Pupil Control | Yes | Yes | Yes | Yes | Yes | Yes |
| $N$ | 160 | 87 | 73 | 160 | 87 | 73 |
| adj. $R^2$ | 0.514 | 0.406 | 0.602 | 0.540 | 0.469 | 0.621 |

*Note:* This table presents the effect of change frame and level frame feedback interacted with students' locus of control when given 1-3 days in advance using a linear regression model including class fixed effects.Column 1 and 4 present the results for the whole sample, columns 2 and 5 present the results for students whose rank improved from exam 1 to exam 2, and columns 3 and 6 present results for students whose rank worsened from exam 1 to exam 2. The dependent variable in columns 1, 2, and is percentage points in exam 3. The dependent variable in columns 4, 5, and 6 is grades in exam 3. (Larger grades are worse grades.) Covariates: percentage points exam 1, percentage points exam 2, gender, own room, foreign name, siblings. Standard errors are reported in parentheses, clustered on classroom level and corrected using biased-reduced linearization. The number of clusters is 10. * p<0.10, ** p<0.05, ***

# Chapter 4

# Salience of Ability Grouping and Biased Belief Formation

## 4.1 Introduction

In recent years, economists have discovered the importance of ability beliefs and social identity for explaining the motivation of individuals to invest in their human capital and to sort into different career paths (Akerlof and Kranton, 2002; Benabou and Tirole, 2002, 2016; Heckman et al., 2006; Dohmen and Falk, 2010, 2011). Whether someone decides to pursue a college degree or to apply for a demanding job depends on how they judge their academic and work-related abilities. In these situations our abilities affect our chances of success and thus our beliefs about them influence the expected payoff of our decisions. Two people with the same abilities may have very different beliefs about them and thus make very different decisions and have very different outcomes in life. While individual characteristics, such as gender (Reuben et al., 2017) and family background (Filippin and Paccagnella, 2012) are known to be correlated with confidence in abilities, the mechanisms bringing about these differences are not well understood. Situational factors, such as the presence of good or bad feedback have been found to influence people's beliefs about their abilities but the effects of more complex social influences, such as the abilities of people in one's immediate environment have only recently attracted the interest of economists.

When people judge their own ability, they may infer their ability level from

comparisons with people in their peer group. For example, someone who finds out that he can do better math than most of his peers may be led to think that he is a good at mathematics and may enjoy it more. However, the person at some point likely encounters another group of people who are on average better at mathematics than he and he might learn that membership in the two groups depends in some way on their mathematics ability. Is it still beneficial for the individual's confidence to be in a weaker group, or not, when both the own as well as other groups can be observed? In other words: Do individuals assign correct weights to ability signals that come from within-group and between-group comparisons?

These questions are important because in different areas of life, such as work and education, groups of different abilities are deliberately formed, often with the intention of improving overall individual performance. However, the empirical evidence suggests that ability grouping may have negative effects on performance (Hanushek and Wößmann, 2006; Malamud and Pop-Eleches, 2011; Guyon et al., 2012; Kerr et al., 2013), although experimental studies that control for environmental factors have found positive effects (Duflo et al., 2011; Booij et al., 2017). More recently, Murphy and Weinhardt (2014) as well as Elsner and Isphording (2017) identified positive effects of having weaker students within one's group on one's long-term academic outcomes and suggest that higher confidence in abilities due to favorable within-group comparisons are the driving force behind this finding. Additionally, experimental studies have shown that between-group comparisons may matter for academic performance. If a person is a member of a group that stereotypically is worse at a given task, salience of this fact may have a negative effect on this person's outcomes (Coffman, 2014; Dee, 2014). In many real-world situations people may have some idea about both their standing within their group and how their group compares to other groups (cf. Trautwein et al., 2006), however the interaction of within-group and between-group information on ability beliefs is not yet well explored. The net effect of assignment to a weaker group on confidence may be negative or positive, depending on the information available to people as well as

how they interpret it.

In this paper, we study the effects of assignment to a weaker group versus a stronger group on confidence and subsequent performance in a laboratory experiment. In our setting, group assignment depends imperfectly on ability so that the ability distributions of the two groups overlap. This implies that the ability signal from group assignment is noisy, which, on the one hand, increases uncertainty that leaves room for interpretation by the subjects and, on the other hand, generates randomness of group assignment that allows for the causal identification of the effect of group assignment on ability beliefs and subsequent performance. We randomly vary whether subjects only receive information about their performance relative to their group or whether they learn additionally whether they were assigned to a weaker or a stronger group and that group assignment depends imperfectly on ability. This allows us to study the causal effects of assignment to a weaker or a stronger group, and its interaction with salience of ability grouping, on confidence in ability and subsequent test outcomes.

We find, first, that the effect of assignment to a weaker group on confidence depends on the salience of ability grouping. When ability grouping is non-salient, it does not matter for subjects' confidence whether they were assigned to the weaker or the stronger group. However, when ability grouping is salient, assignment to the weaker group makes people less confident in their abilities. Second, subjects on average gave quite correct estimates of their ability rank, when grouping was non-salient. However, when grouping was salient, subjects who were assigned to the stronger group were significantly overconfident while subjects who were assigned to the weaker group were significantly underconfident, indicating that people *overweighed* ability signals coming from *between-group* information. Also, subjects who learned they were assigned to the weaker group were more underconfident than subjects who learned they were assigned to the stronger group were overconfident. This difference cannot be explained by lower ability subjects reporting less correct beliefs, rather, it shows that people *overweighed negatively surprising* information as com-

pared to positively surprising information. Third, results also suggest that higher ability subjects perform worse if they learn they were assigned to a weaker group, while lower ability subjects perform better when learning that they were assigned to a weaker group. We do not find this difference when ability grouping is non-salient. These findings indicate that when people are sorted into different ability groups, within-group and between-group information interact in complex ways to affect ability beliefs and subsequent performance.

To our knowledge, this is the first study to show causal effects of ability grouping on ability beliefs. It shows that both within-group and between-group information, which may not be processed symmetrically, matter for people's beliefs about their abilities. The results of this study demonstrate that the effects of one's group's abilities on beliefs in own ability and subsequent performance are sensitive to information about the group assignment process. For this reason, one should be careful when interpreting effects of peer group ability on performance from field experiments in which the rules determining group assignment are non-salient (as e.g. in Duflo et al. 2011; Carrell et al. 2013; Booij et al. 2017) as these effects may not hold once people understand that groups of different abilities were formed deliberately.

The paper is structured as follows: Section 2 summarizes the related literature, Section 3 describes the experimental design, Section 4 presents and discusses the results and Section 5 concludes.

## 4.2 Related Literature

Higher confidence in one's abilities has been found to have beneficial effects on one's educational and labor market outcomes (Heckman et al., 2006; Cebi, 2007; Heineck and Anger, 2010). Recent evidence also suggests that confidence in one's abilities may be influenced by the abilities of people in one's peer group. Murphy and Weinhardt (2014) find that, controlling for own ability as measured by standardized test scores at age 11, an increase in rank during one's primary school class has a large and significant positive effect on test scores at age 14. The authors also find

that the development of subject-specific confidence is the most likely driver of these effects. Similarly, Elsner and Isphording (2017) find that, controlling for own ability, students who have a higher rank within their cohort in high school perceive their intelligence to be higher, have higher expectations about their future careers and are more likely to go to college and complete a degree. These studies run counter to the received wisdom from the peer effects literature that better peers are better for academic performance but provide evidence in favor of the so called "big-fish-little-pond effect" (Marsh, 1987), a popular proposition claiming that assignment to a peer group with lower skills increases one's confidence in ability[1] that is based on theories of social comparison processes (Festinger, 1954).

On the contrary, the experimental literature highlights the importance of between-group comparisons. For example, people infer individual characteristics from group characteristics, which may lead to self-stereotyping (Coffman, 2014; Dee, 2014). While the traditional economic approach assumed that people form rational expectations about a group member in terms of the aggregate distribution of the characteristics of his group (e.g. Phelps 1972; Arrow 1973, for an overview of the literature see Fang and Moro, 2011), the social cognition approach, which has influenced behavioral economics, holds that people form intuitive generalizations that allow them to save mental resources but which may lead to biases in beliefs. The generalizations are based on real differences between groups and as such contain a "kernel of truth" but they are selective and may exaggerate between-group differences while tending to underweigh within-group differences (Schneider, 2004). Several studies have provided evidence in support of this hypothesis. Recently, Dee (2014) presents empirical evidence from a framed field experiment that self-stereotyping effects can

[1]Trautwein et al. (2006) qualify this statement based on correlations between confidence in mathematics ability and mathematics test scores of students in German secondary schools. In their study, schools are either in the high, middle, or low ability track or comprehensive schools that incorporate all three tracks. Controlling for math ability, within tracked schools, students' confidence is higher in schools of lower ability tracks. However, in comprehensive schools where different ability tracks can be found under the same roof, making ability tracking highly observable for students every day, controlling for ability, students' confidence in the higher and the lower tracks did not differ significantly. These observations support the central assumption of this study that both within-group and between-group comparisons of abilities as well as the salience of ability tracking should matter for students' confidence in their abilities.

be relevant in an education context: Students at a selective college were randomly assigned to a treatment that primed their awareness of a negatively stereotyped identity (here: a student-athlete). This social-identity manipulation reduced the test performance of athletes relative to non-athletes in spite of causing them to attempt to answer more questions. Similarly, Coffman (2014) finds that, conditional on measured ability, individuals are less willing to contribute ideas in areas that are stereotypically outside of their gender domain, which is largely driven by self-assessments rather than by fear of discrimination, and cannot be easily corrected by providing contrary feedback. Furthermore, Albrecht et al. (2013) show that individuals from groups that perform badly on average receive low evaluations, even when it is known that the individuals themselves perform well. This shows that people incorporate group information when evaluating individuals even in cases where it is irrelevant. However, Fryer et al. (2008) cannot reproduce the standard finding that female performance declines in absolute terms when the experimental instructions include a passage emphasizing that men outperform women on a given test.

There is a trade-off between a favorable within-group comparison and a favorable between-group comparison of abilities as the the within-group effect ("big-fish-little-pond-effect") runs counter to the between-group effect, also called the effect of "stereotype threat": One can either be "a bigger fish in a smaller pond" or "a smaller fish in a bigger pond" and it is not ex-ante clear which is better for confidence in abilities. When assigning correct weights to within-group and between-group ability signals, it should not matter for one's confidence whether one is assigned to the weaker or the stronger group as between-group information would counterbalance within-group information. However, subjects could possibly place a greater weight on within-group or between-group information, on positive (Eil and Rao, 2011; Mobius et al., 2011; Grossman and Owens, 2012; Wiswall and Zafar, 2015) or negative (Ertac, 2011) information, or exhibit other forms of biased belief formation (see e.g. Albrecht et al. 2013; Butler 2016).

Furthermore, negative information about one's abilities could both induce higher

(Kuhnen and Tymula, 2012; Azmat et al., 2016; Fischer and Wagner, Chapter 3 of this thesis) or lower (Buser, 2016) subsequent performance, depending on how subjects' effort depends on their ability beliefs. Overall, there is mixed evidence on the association between feedback and performance (Kluger and DeNisi, 1998; Hattie and Timperley, 2007), possibly because the relationship between ability beliefs and effort is complex. In recent years, a number of studies has highlighted the importance of distinguishing between confidence in abilities that are complements and confidence in abilities that are substitutes to effort (Santos-Pinto 2008; Ederer 2010; Caliendo et al. 2015; Spinnewijn 2015; Fischer and Sliwka in Chapter 5 of this thesis). In a setting of human capital investment, Fischer and Sliwka (Chapter 5) distinguish between confidence in learning ability – the belief that one can raise one's probability of being successful by exerting effort – and confidence in prior knowledge – the belief that one's probability of being successful is already high prior to investing any additional effort. The authors show experimentally that the use of feedback that raises confidence in learning ability increases motivation to strive towards a better outcome. However, the use of feedback that raises confidence in prior knowledge decreases motivation to strive towards a better outcome for individuals for whom success was more likely at baseline. Fischer and Sliwka's notion of confidence in the effectiveness of effort is equivalent to Benabou and Tirole's (2002, 2003) notion of confidence as an agent's (rational) belief in her own marginal product of effort and possibly captures ability beliefs *positively* related to educational outcomes, as e.g. in Heckman et al. (2006), Cebi (2007), Heineck and Anger (2010), Murphy and Weinhardt (2014), and Elsner and Isphording (2017). In contrast, their notion of confidence in the baseline probability of success possibly describes the kind of belief measured in studies that find higher confidence to have *negative* effects on people's outcomes (see e.g. Camerer and Lovallo (1999), Malmendier and Tate (2005), and Niederle and Vesterlund (2007)). The current study uses within- and between-group information to manipulate people's confidence in their learning ability which, according to theory, is complementary to effort. We therefore expect feedback that

bolsters this ability belief to positively influence effort and in turn performance.

## 4.3 Experimental Procedure

The experiment was conducted in November 2016 at the Cologne Laboratory of Economic Research[2] using the experimental software z-tree (Fischbacher, 2007). Participants were recruited using the software ORSEE (Greiner, 2004) and, upon arrival, were randomly assigned to one of 32 terminals that were divided by panels. Before the experiment started, participants received instructions that communication with each other and the use of mobile phones or pens was not permitted and that compliance with this rule would be monitored during the whole experiment. These, and all of the following instructions were given on-screen. Participants were also informed that they would receive 4 euros for participating in the experiment and that they could earn additional money by correctly answering questions in several "learning tests". They then received instructions for the first learning test, including the task and the reward scheme, and had to correctly solve a sample question before they could proceed to the test.

**First test** Each task consisted in assigning to a city name the first digit of its corresponding four digit fictitious "city code". The test consisted of 36 tasks and subjects earned 0.10 euros for each correctly solved task. Before the test phase, there was a 10 minutes learning phase during which subjects could study the city name and code pairs. As shown in Figure 4.7 in Appendix 4.7.1, during the learning phase city names were listed alphabetically in three columns and the corresponding city codes were displayed next to them for three seconds when the button with the respective name was pressed.[3] Subjects could press these buttons as often as they wanted, without incurring any costs, and in quick succession such that several codes were be displayed at once. Subjects who did not want to study could leave the study

[2]Financial support of the Deutsche Forschungsgemeinschaft (DFG) through grant FOR1371 is gratefully acknowledged.

[3]This feature was meant to capture subjects' intensity of learning.

screen and spend time looking at comics but could return to studying at any time without this having any implications for them beyond the loss of time. This element was introduced to allow for opportunity costs of studying. During the 6 minutes test phase (see Figure 4.8 in Appendix 4.7.1), city names were displayed in random order and the correct digit had to be filled in next to them.

**Feedback stage (treatment randomization)** After the first test, subjects were informed that they would receive feedback about their "learning ability" relative to the other participants based on their result in the learning test. On the next screen, subjects received their feedback. The assignment mechanism of the feedback was as follows: Subjects were randomly assigned to one type of "ability grouping system", which was either "Non-salient Grouping" or "Salient Grouping". Next, the experimental software assigned each subject either to the "Stronger Learners" or the "Weaker Learners" group. Here, the probability of assignment differed depending on a person's performance in the first test. Those who in the first test performed in the upper half relative to the other participants in the session (percentile rank $<0.5$ relative to all) were assigned to "Stronger Learners" with a probability of 2/3 and were assigned to "Weaker Learners" with a probability of 1/3. On the contrary, those who in the first test performed in the lower half (percentile rank $>0.5$ relative to all) were assigned to "Stronger Learners" with a probability of 1/3 and were assigned to "Weaker Learners" with a probability of 2/3. Depending on the group someone was assigned to, the experimental software then computed a person's rank within her group and determined whether this rank was in the upper (percentile rank $<0.5$ relative to group) or the lower half (percentile rank $>0.5$ relative to group).

As summarized in Table 4.1 subjects received different information, depending on the treatment (i.e. "ability grouping system") they were assigned to (The messages displayed to subjects in each treatment can be found in Table 4.7.1 in Appendix 4.7.1.):

Non-salient Grouping: Subjects received feedback relative to their group, which

they knew was half the session's participants and did not learn anything about the characteristics of the group.

Salient Grouping: Subjects received both feedback relative to their group and, on the same screen, they also received information about whether they were assigned to the "Stronger Learners" or the "Weaker Learners" group, which they knew consisted of half the session's participants. They also knew that their assignment depended imperfectly on their ability as they were told that "a better result makes it much more likely to be assigned to the stronger learners". Table 4.1 summarizes the information provided in each treatment.

Table 4.1: Information by Treatment

| *Treatment:* | **Non-salient grouping** | **Salient grouping** |
|---|---|---|
| *Information:* | upper/lower half in group | upper/lower half in group + stronger/weaker group |

**Belief elicitation** After receiving feedback, subjects were asked to estimate their rank with respect to their performance and their effort (in terms of clicks on city names in the learning phase) in the first test relative to the other participants in the room (session). They knew that for each of the two rank estimates they would earn one euro if it was correct.

**Second test and questionnaire** After indicating their beliefs the next screen informed subjects that the second test was of the same type, length and duration as the first test but that this time they would earn 0.20 euros (as compared to 0.10 euros in the first test) for each correctly solved task. They were also informed that, unlike after the first test, they would not be able to earn any money by estimating their performance or effort rank relative to other participants. After having read these instructions subjects proceeded to the learning stage of the second test. As can be seen in Figure 4.12 and Figure 4.13 in Appendix 4.7.1, the second test looked identical to the first test, it only contained other city names and numbers. When the test was designed, the questions were randomly assigned to test 1 and test 2 in

order to create "parallel" tests of the same difficulty. After the second test, subjects were asked to indicate in which of the two tests they believed they performed better and in which they had invested more effort. They could earn 0.50 euros for each correct answer. They then filled in a short demographic survey and learned their earnings from each stage of the experiment.

## 4.4 Experimental Results

The experiment lasted approximately 50 minutes and participants on average earned 11.41 euros. In total 7 sessions were conducted, which were orthogonal to treatments to rule out self-selection. All participants were university students, who were on average in their 6th semester of study. 49 percent of participants were female. On average, 19.8 out of 36 questions in the first test and 22.7 out of 36 questions in the second test were answered correctly. There were 79 participants in the non-salient grouping treatment and 78 participants in the salient grouping treatment.[4]

In Section 4.4.1 we will analyze the effect of salience of ability grouping and assigned group on confidence. Separately for salient and non-salient grouping, we will then explore the response of people with higher and lower ability to higher and lower group assignment. In Section 4.4.2 we will then shed light on the mechanisms underlying the observed results. In particular, we will (i) address the question to what extended the observed responses are rational given the information provided to people and (ii) investigate whether information processing is affected differently by positive and negative within-group and between-group information. In order to do so, we will derive rank predictions conditional on feedback and will then study how well different groups match their predicted ranks. Finally, in Section 4.4.3 we

[4]A treatment where participants were not assigned to a group was also conducted to check whether these two treatments lead to an overall distortion of beliefs. In this benchmark treatment subjects received feedback about whether their performance was in the upper or lower half relative to the whole session. 63 subjects originally participated in this treatment, however only 36 observations are usable due to a programming error. This error affected participants randomly, so that this treatment is still completely balanced to the other two treatments, as can be seen in Table 4.6 in Appendix 4.7.2. It may thus, as intended, serve to benchmark the distortions caused by the two treatments relevant to our research question.

will analyze the effects of group assignment and salience of ability grouping on test outcomes.

We expect, based on prior research (Murphy and Weinhardt, 2014; Elsner and Isphording, 2017), that when subjects only learn about their standing within their group, they become more confident when they are assigned to the weaker group. However, when learning about both their standing within their group and their group's standing relative to another group, this effect disappears if subjects assign correct weights to within-group and between-group ability signals, as in this case between-group information counterbalances within-group information (cf. Trautwein et al., 2006). However, if subjects overweigh between-group information, the effect of weaker group assignment is negative, while if they overweigh within-group information the effect of weaker group assignment is still positive. Furthermore, the current study gives people feedback about their "learning ability" in order to influence people's beliefs in the marginal productivity of learning effort, which according to theory (e.g. Fischer and Sliwka in Chapter 5 of this thesis), is positively related to learning effort. We therefore expect feedback that strengthens this ability belief to positively influence performance.

### 4.4.1 Effects of Salience of Ability Grouping and Group Assignment on Confidence

Our first variable of interest is confidence, which we define as

$$Confidence = Rank - RankBelief.$$

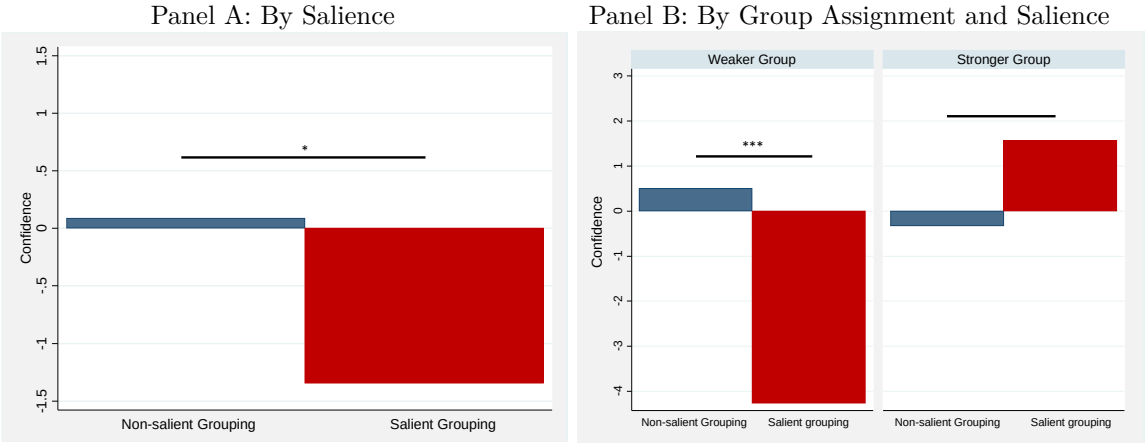
Recall from Section 4.3 that we elicited the $RankBelief$ by asking subjects to estimate their rank in the first test relative to all other participants in their session. Likewise, the $Rank$ measures a subject's actual performance in the first test relative to all other participants in the same session. Thus, our confidence measure is very intuitive as it captures the degree to which subjects overestimate or underestimate

their performance relative to the other participants: If someone overestimates his performance relative to the other participants he will have $Confidence > 0$, while if he underestimates his performance relative to the other participants he will have $Confidence < 0$.

In the following, we will study the causal effects of salience of ability grouping as well as its interaction with assigned group on confidence. Then, we will study these two effects, as well as the overall effect of group assignment, separately for higher and lower ability subjects. Note that while the causal effect of salience as well as its interaction with group assignment can be studied for the whole sample, the causal effect of group assignment by itself has to be studied separately for the higher and lower ability subjects as these two groups had different assignment probabilities.[5]

[5]Higher ability subjects (who performed above the median in the first test) had a probability of being assigned to the stronger group that was twice as large as the probability of the lower ability subjects (who performed below the median in the first test). This means that, as intended, individuals in the weaker group had on average lower ability than individuals in the stronger group. Our confidence measure captures ability beliefs relative to true ability, so group differences in ability are controlled for in the graphs. However, as subjects had to state their beliefs in terms of ranks (#ranks = #subjects in session), the belief scale is restricted from above and from below, which means that higher ability subjects are more restricted in their possibility to report overconfidence than in their possibility to report underconfidence, while lower ability subjects are more restricted in their possibility to report underconfidence than in their possibility to report overconfidence. This may induce the overconfidence of higher ability subjects and the underconfidence of lower ability subjects to be underestimated. Within these two groups, the probability of being assigned to any of the two groups was perfectly random so that the restriction with respect to reporting overconfidence and underconfidence affected people assigned to the stronger group and the weaker group equally. Hence, by analyzing the effects of group assignment separately for higher and lower ability subjects, we can identify the causal effects of assignment to the weaker or stronger group on confidence.

Figure 4.1: Effects of Salience of Ability Grouping and Group Assignment on Confidence




*Note*: Panel A shows the effect of salience of ability grouping on confidence. Panel B shows the interaction effect of salience of ability grouping and group assignment on confidence.

As can be seen in Panel A of Figure 4.1, confidence was higher when ability grouping was non-salient than when ability grouping was salient. Man-Whitney U (M-W U) tests[6] show that this difference is weakly significant. As can be seen in Panel B of Figure 4.1 subjects who were assigned to the weaker group but did not know that their group was the weaker one were more confident than subjects who were assigned to the weaker group and knew that their group was the weaker one (M-W U test: p=0.00). On the contrary, when one was assigned to the stronger group, knowing whether one's group was the stronger one did not significantly affect one's confidence (M-W U test: p=0.32).

[6]All tests in this paper are two-sided, unless stated otherwise.

Table 4.2: Effects of Salience of Ability Grouping and Group Assignment on Confidence

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable: Confidence | If Lower A. | If Higher A. | If Lower A. | If Higher A. |
| Non-salient Grouping | 0.949 | 3.349** | 3.952** | 10.10*** |
| | (0.55) | (2.54) | (2.07) | (5.02) |
| Stronger Group | 3.282* | 4.191*** | 6.839*** | 10.65*** |
| | (1.78) | (3.25) | (2.71) | (6.36) |
| Non-salient Gr. × Stronger Group | | | -7.482* | -11.07*** |
| | | | (-1.97) | (-4.33) |
| Observations | 76 | 81 | 76 | 81 |
| $R^2$ | 0.117 | 0.242 | 0.173 | 0.413 |

*Note*: This table presents the effect of non-salient versus salient ability grouping and assignment to a stronger versus a weaker group using a linear regression model including a constant, session fixed effects and robust standard errors. Dependent variable: confidence. Columns 1 and 3 (2 and 4) show results for lower (higher) ability subjects. t statistics are reported in parentheses * p<0.10, ** p<0.05, *** p<0.01.

In Table 4.2 we analyze, separately for higher and lower ability individuals, the effects of salience, of group assignment, as well as of the interaction between the two. The regressions are estimated by ordinary least squares and contain heteroscedasticity robust standard errors as well as session dummies and a constant, but no other control variables. Thus, all the coefficients show causal effects of our intervention. As can be seen in Columns 1 and 2 both lower and higher ability subjects were on average more confident (by 3.3 and 4.2 ranks, respectively) if they were assigned to the stronger group. These effects are marginally and highly significant, respectively. However, only higher ability subjects are affected by the salience of ability grouping overall. Thus, the difference in confidence shown in Panel A of Figure 4.1 are largely driven by the response of higher ability subjects. They were on average 3.3 ranks more confident when ability grouping was non-salient. Columns 3 and 4 present results for the interaction effects between group assignment and salience of the assignment mechanism. Qualitatively, lower and higher ability subjects respond similarly but the effects seem to be larger for higher ability subjects. When ability grouping is salient, both lower and higher ability subjects are more confident when they are assigned to the stronger group (by 6.8 and 10.7 ranks, respectively). Both effects are highly significant. Those who were assigned to the weaker group were

more confident (by 4.0 and 10.1 rank, respectively) when they did not learn that their group was the weaker one. These effects are significant at the 5% and the 1% level, respectively. F-tests show that when ability grouping was non-salient, it did not matter for lower or higher ability subjects whether they were assigned to the weaker or the stronger group (for both p=0.81). Hence the differences presented in Panel B of Figure 4.1 are driven by both lower and higher ability subjects.[7]

### 4.4.2 Mechanisms

The above results show that when group assignment is salient, assignment to the weaker group causes individuals to be less confident than assignment to the stronger group. Furthermore, weaker group assignment causes subjects to be less confident when grouping is salient than when grouping is non-salient. The mechanisms underlying these observations can be further explored on three levels. First, we can investigate to what extend salient and non-salient ability grouping leads to a *de-calibration* of beliefs, i.e. to what extend they make people overconfident or underconfident.[8] Second, we can explore to what extend non-salient and salient ability grouping lead people to state "*irrational*"[9] beliefs, i.e. rank beliefs that could not possibly be correct given the feedback someone received. Third, we can shed light on how non-salient and salient ability grouping affect the *distributions* of beliefs. This may help us to better understand the average treatment effects as well as the

[7]Interestingly, higher but not lower ability subjects' beliefs in their intensity of effort (in terms of clicks), when ability grouping was salient, responds significantly to group assignment: When learning they were assigned to the weaker group, higher ability subjects believe to have exerted less effort than when learning they were assigned to the stronger group. This may indicate that higher ability subjects attribute weaker group assignment more strongly to effort (rather than to ability) than lower ability subjects.

[8]In the benchmark treatment, in which people were not assigned to different groups and received feedback relative to the whole session, people's mean confidence was 0.31, which is not significantly different from 0 (t-test: p=0.69). Hence, without ability grouping, subjects were on average well calibrated.

[9]The feedback given to each person implied that there were certain ranks they were definitely not occupying. As people were paid for correct rank estimates, it was never optimal for one's monetary payoff to report rank beliefs that are definitely false. However, one could think of a model where an agent benefits from incorrect beliefs, e.g. with respect to his self-image or his motivation. In this case, false beliefs could possibly be optimal. In our setting, we will abstract from this possibility and will call beliefs "irrational" if they indicate a rank that was impossible for a person given the information they had received.

decalibration of beliefs.

**Overconfidence and underconfidence** When grouping is salient, confidence could be lower with weaker group assignment than with stronger group assignment due to (1) weaker group assignment making people underconfident and/or (2) stronger group assignment making them overconfident. Panel B of Figure 4.1 suggests that the effect is driven mostly by salient grouping making people assigned to the weaker group underconfident, while they seem well calibrated when grouping is non-salient. Furthermore, when ability grouping is salient, people tend to be on average less confident than when ability grouping is non-salient. This could be due to (1) non-salient grouping making people overconfident and/or (2) salient grouping making people underconfident. Panel A of Figure 4.1 suggests that while with non-salient grouping people have on average quite correct beliefs, they seem to be very underconfident on average with salient ability grouping.

Using one-sided t-tests of the means of the four groups (stronger group–non-salient / weaker group–non-salient / stronger group–salient / weaker group–salient) in Panel B of Figure 4.1 against the null hypothesis that people have correct beliefs (Confidence=0) reveals that when grouping was non-salient, subjects were neither significantly overconfident when assigned to the weaker group (p=0.30) nor significantly underconfident when assigned to the stronger group (p=0.36). However, if grouping was salient, subjects who were assigned to the weaker group were significantly underconfident (p=0.00) and subjects who were assigned to the stronger group were weakly significantly overconfident (p=0.08). Furthermore, a M-W U test reveals that if grouping was salient, people who were assigned to the weaker group were significantly more underconfident than people who were assigned to the stronger group were overconfident (p=0.03). This shows that people assigned a larger weight to the ability signal from group assignment when it was negative than if it was positive.

Overall, people become more decalibrated by salient than by non-salient ability grouping. When ability grouping is salient, they are more decalibrated if they are

assigned to the weaker group than if they are assigned to the stronger group.

**"Irrational" beliefs** In the following, we will address the question to what extend the stronger decalibration from salient than from non-salient grouping is "irrational" given the feedback information subjects received. The feedback given to each person, while imprecise about their relative position, ruled out certain ranks for them.Thus some rank beliefs were "irrational" for them to hold. We will also shed light on the mechanisms that may explain why, when ability grouping is salient, weaker group assignment leads people to become more decalibrated than stronger group assignment.

Note that the feedback types explained in Section 4.3 are not equal to the four groups analyzed above (stronger group–non-salient / weaker group–non-salient / stronger group–salient / weaker group–salient). This is because in the non-salient grouping treatment (stronger group–non-salient / weaker group–non-salient) subjects did not learn their group assignment but only which half they occupied within their group. Hence, the two feedback types with non-salient grouping are "upper half within group" and "lower half within group", which we will call "Non-salient Grouping – 1" and "Non-salient Grouping – 2", respectively. By contrast, in the salient grouping treatment, people learned both whether their group was the weaker or the stronger one as well as their half within their group. Thus, with salient grouping, we have four feedback types: "upper half in stronger group", "lower half in stronger group", "upper half in weaker group" and "lower half in weaker group", which we will call "Salient Grouping – 1", "Salient Grouping – 2", "Salient Grouping – 3", and "Salient Grouping – 4", respectively. Furthermore, subjects knew that their group assignment depended imperfectly on their ability. Hence, they knew that stronger group assignment did not necessarily imply that one's performance was above average, while weaker group assignment did not necessarily imply that one's performance was below average.

Figure 4.2: Information Content of Feedback and Distribution of Beliefs

| Treatment | Feedback | | Information/Belief Distribution | | | |
|---|---|---|---|---|---|---|
| | Group | Half within Group | 1st Quart. (0.00-0.25) | 2nd Quart. (0.26-0.50) | 3rd Quart. (0.51.-0.75) | 4th Quart. (0.76-1.00) |
| Non-salient Grouping – 1 | ? | Upper | 60.47 | 37.20 | 0 | 2.33 |
| Non-salient Grouping – 2 | ? | Lower | 2.78 | 2.78 | 38.88 | 61.11 |
| Salient Grouping – 1 | Stronger | Upper | 72.22 | 27.78 | 0 | 0 |
| Salient Grouping – 2 | Stronger | Lower | 0 | 52.38 | 14.29 | 33.33 |
| Salient Grouping – 3 | Weaker | Upper | 5.56 | 0 | 77.77 | 16.67 |
| Salient Grouping – 4 | Weaker | Lower | 0 | 9.52 | 9.52 | 80.95 |

A, B, C

Likely Possible Impossible

*Note*: This table indicates the likelihood, conditional on feedback, of being ranked in a given quartile (dark gray: likely, light gray: possible, white: impossible). The numbers indicate the percentage of people believing to be ranked in a given quartile, conditional on feedback.

Figure 4.2 shows the six different types of feedback that were given during the experiment. For example, if someone was in the non-salient ability grouping treatment he was either told that he was in the upper half within his group or that he was in the lower half within this group. If he was in the upper half within his group (feedback type "Non-salient Grouping – 1"), and the ability distributions of the two groups were not too different, he was likely in the upper half with respect to all the participants in the session. However, it was theoretically possible that his group was much worse than the other group. In this case, being in the upper half within this group could possibly entail being only in the $3^{rd}$ quartile with respect to all participants. However, even if his group was so bad compared to the other group that the two groups' ability rank distributions did not overlap, given that he was told he was in the upper half within his group, it was impossible that he occupied an ability rank in the $4^{th}$ quartile (percentile ranks $>= 0.75$) with respect to all people in the session. Applying the same reasoning to the other five types of feedback as well produces the different zones (likely range, possible (less likely) range, impossible range) that are indicated by the different shadings for the four quartiles. The numbers in Table 4.2 indicate the percentage of people, in a given feedback category, who reported a rank belief in the respective quartile. To give

an applied example, consider subjects who were in salient grouping and were told that they occupied a rank in the upper half of the stronger group (feedback type "Salient Grouping – 1). Among them 72.22 percent indicated a rank belief in the first quartile (for them, the likely range), while 27.78 percent reported a rank belief in the second quartile (for them, the possible range). None of these people reported a rank in the $3^{rd}$ or $4^{th}$ quartile. We can conclude that none of the people who received this type of feedback reported an "irrational" belief.

With non-salient grouping, the two groups ("upper half within group" and "lower half within group") have similar belief distributions over the likely, possible and impossible range. However, with salient grouping, the picture is different. Here, of those who were assigned to the weaker group 16.7 and 9.5 percent, respectively, report beliefs in the impossible range while none of those assigned to the stronger group do so. Furthermore, those who were assigned to the weaker group seem to state fewer beliefs in the possible range than those assigned to the stronger group. Among those of the weaker group, the proportion of people stating a belief in the likely range seems to be larger (at 77.77 and 80.95 percent, respectively) than among those of the stronger group (52.38 and 72.22 percent, respectively). In the following, we will study how similar, overall and within the four quartiles, the belief distributions of people who received the different types of feedback are.

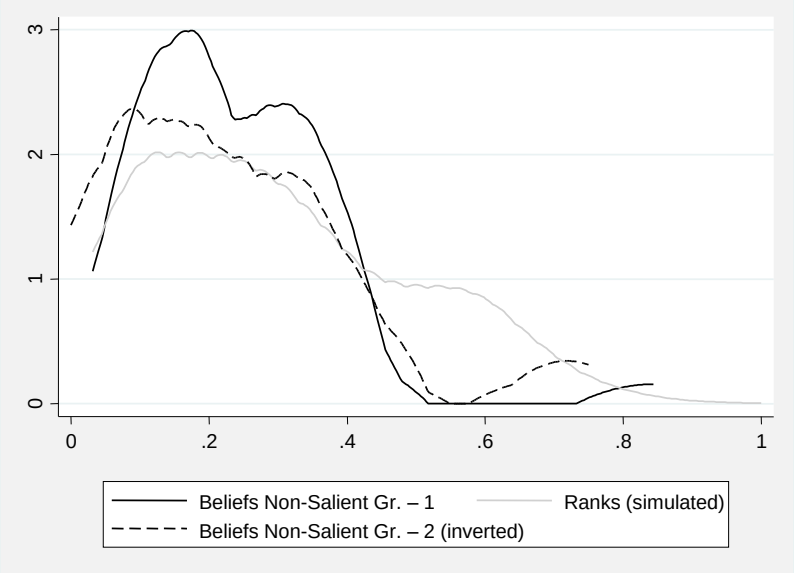
**Belief distributions by feedback types** As as shown in Table 4.2, people who received the different feedback types had different ranges of likely, possible, and impossible beliefs. To illustrate this, Figure 4.14 in Appendix 4.7.3 shows the expected ability rank distributions by feedback type resulting from our assignment mechanism. We can see that the expected rank distributions for subjects who received feedback types "Non-salient Grouping – 1" and "Non-salient Grouping – 2", and likewise for "Salient Grouping – 1" and "Salient Grouping – 4" as well as for "Salient Grouping – 2" and "Salient Grouping – 3" are mirror images of each other. Hence, within these pairs of feedback types the rank distributions that subjects had to match with their beliefs were the same except for being inverted. Thus, a

straightforward way for testing whether the belief distributions differed from each other, conditional on feedback, within each of the three pairs is to invert the elicited rank beliefs of one of the groups within each of the pairs. Next, we can run statistical tests for the equality of distributions.

Figures 4.3, 4.4, and 4.5 show the inverted belief distributions from the feedback types whose expected rank distributions are shown on the right hand side of Figure 4.14 in Appendix 4.7.2 mapped onto the belief distributions from the feedback types whose expected rank distributions are shown on the left hand side. Furthermore, they are depicting the expected rank distributions that are shown in Figure 4.14, which are identical within each pair after the right hand side distributions have been inverted. As can be seen in Figure 4.3, which corresponds to comparison "A" in Figure 4.2 and Panel A in Figure 4.14, with non-salient grouping, when people receive positive feedback (Non-salient Grouping – 1 (NSG–1)), they have a very similar belief distribution, conditional on feedback, as people who receive negative feedback (Non-salient Grouping – 2 (NSG–2)). Subjects in NSG–1 seem largely not to take into consideration that they could occupy a rank in the lower half while subjects in NSG–2 seem to largely ignore their rank could be in the upper half with respect to the whole session. A Kolmogorov–Smirnov test shows that the two distributions are not significantly different overall (p=0.58)[10]. Testing the distributions in the four quartiles separately reveals that the $1^{st}$ quartile of NSG–1 is not significantly different from the $4^{th}$ quartile of NSG–4 and the $2^{nd}$ quartile of NSG–1 is not significantly different from the $3^{rd}$ quartile of NSG–2. However, while NSG–1 has no observations in the $3^{rd}$ and the $4^{th}$ quartile NSG–2 has observations both in the $2^{nd}$ and in the $1^{st}$ quartile (as can also be seen in Figure 4.2) .

[10]For all Kolmogorov–Smirnov tests in the paper exact p-values from combined (two-sided) tests are reported

Figure 4.3: Comparison of Beliefs in Non-Transparent grouping (A)

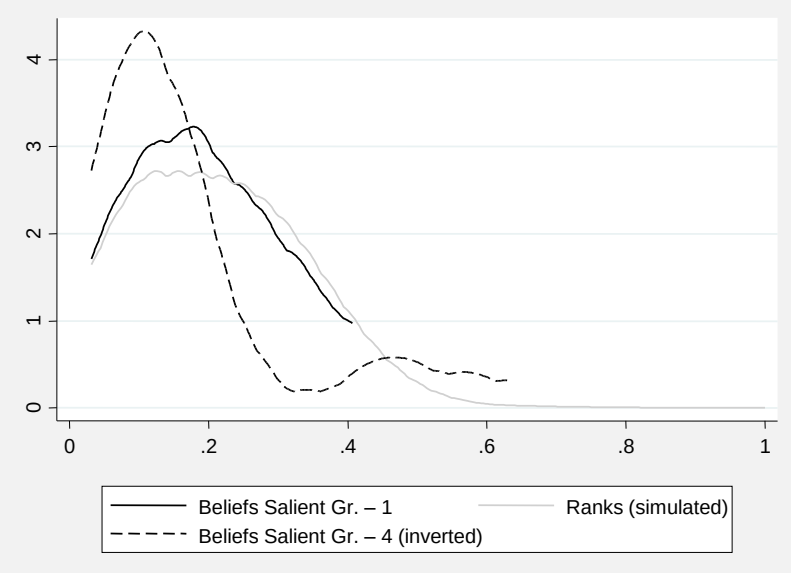



*Note*: This graph shows the distributions of the rank beliefs of subjects who received feedback type "Non-Salient Grouping – 1", the inverted rank beliefs of subjects who received feedback type "Non-Salient Grouping – 2", and the (inverted) expected ability rank distribution for subjects who received feedback type "non-Salient Grouping – 1 " ("Non-Salient Grouping – 2 ").

As can be seen in Figure 4.4, which corresponds to comparison "B" in Figure 4.2 and Panel B in Figure 4.14, with salient grouping, when people get extreme feedback ("upper half in stronger group" or "lower half in weaker group") and it is negative (Salient Grouping – 4 (SG–4)), "lower half in weaker group"), they tend to interpret it more extremely than when they get positive feedback (Salient Grouping – 1 (SG–1), "upper half in stronger group"). However, among those who get negative feedback some take into account the possibility that they might have been in the upper half overall (inverted percentile rank >0.5). When people get extreme positive feedback they seem to have surprisingly correct beliefs overall. However, they seem to ignore the possibility that they might have performed in the lower half (percentile rank >0.5). Kolmogorov–Smirnov tests show that the two distributions are not significantly different overall (p=0.25). Testing the four quartiles separately reveals that the $2^{nd}$ quartile in SG – 1 and the $3^{rd}$ quartile in SG – 4 ((inverted) percentile rank >0.25 and <0.5) are weakly significantly different from each other (p=0.09).

Furthermore, the distributions are different in the $3^{rd}$ quartile in SG – 1 and the $2^{nd}$ quartile in SG – 4 ((inverted) percentile rank >0.50 and <0.75), as SG–1 does not have any observations in the $3^{rd}$ quartile while SG–4 has observations in the $2^{nd}$ quartile (as can also be seen in Figure 4.2).

Figure 4.4: Comparison of Beliefs in Transparent grouping (Extreme Feedback) (B)

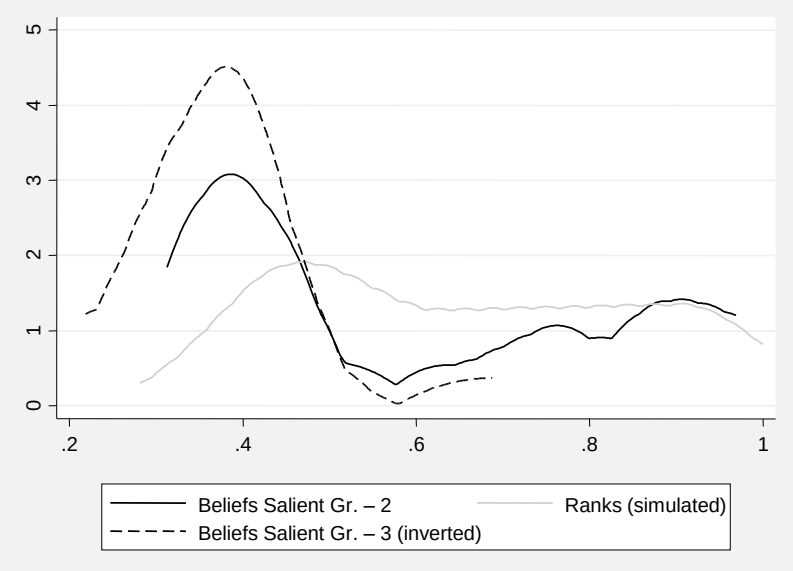



*Note*: This graph shows the distributions of the rank beliefs of subjects who received feedback type "Salient Grouping – 1", the inverted rank beliefs of subjects who received feedback type "Salient Grouping – 4", and the (inverted) expected ability rank distribution for subjects who received feedback type "Salient Grouping – 1" ("Salient Grouping – 4").

As can be seen in Figure 4.5, which corresponds to comparison "C" in Figure 4.2 and Panel C in Figure 4.14, with salient grouping, when people get positive feedback about their group but negative feedback about their standing within their group (Salient-grouping – 2 (SG–2), "lower half in stronger group"), many of them correctly take into account that they might in fact have performed in the lower half relative to the whole session. However, when people get negative feedback about their group but positive feedback about their standing within their group (Salient-grouping – 3 (SG–3), "upper half in weaker group"), they largely ignore the possibility that they might have performed in the upper half overall. Kolmogorov–Smirnov tests show that the belief distributions with these two ambivalent feedback types

are significantly different overall (p=0.039). Testing the four quartiles separately reveals that the $2^{nd}$ quartile of SG–2 and the $3^{rd}$ quartile of SG–3 as well as the $3^{rd}$ quartile of SG–2 and the $2^{nd}$ quartile of SG–3 are not significantly different from each other. The $1^{st}$ quartile of SG–2 has no observations while the $4^{th}$ quartile of SG–3 does. Furthermore, the $4^{th}$ quartile of SG–2 does have observations while the $1^{st}$ quartile of SG–3 does not (as can also be seen in Figure 4.2). Those who received "lower half in stronger group" feedback seem to correctly take into account that the partial randomness of our group assignment mechanism implies that one may have below average performance in spite of being assigned to the stronger group. On the contrary, those who received "upper half in weaker group" feedback seem to ignore the partial randomness of our group assignment mechanism and that they may well have above average performance in spite of being assigned to the weaker group. Note that the group who seems to ignore the partial randomness of assignment has on average higher performance in the first test than the group who takes it into account (M-W U: p=0.062). Thus, the resulting more decalibrated beliefs among those receiving bad between-group and good within-group information than those receiving good between-group and bad within-group information cannot be explained by the former having lower ability as measured by the test (which may be correlated with the ability to understand the feedback). Rather, negatively surprising group assignment seems to lead to a larger decalibration of beliefs than positively surprising group assignment.

Figure 4.5: Comparison of Beliefs in Transparent grouping (Ambivalent Feedback) (C)



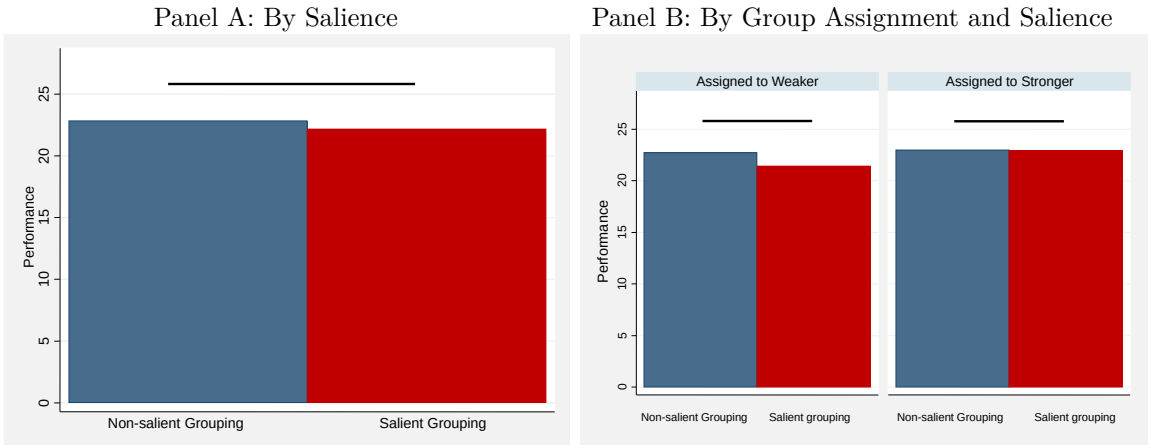

*Note*: This graph shows the distributions of the rank beliefs of subjects who received feedback type "Salient Grouping – 2", the inverted rank beliefs of subjects who received feedback type "Salient Grouping – 3", and the (inverted) expected ability rank distribution for subjects who received feedback type "Salient Grouping – 2" ("Salient Grouping – 2").

Implications of these findings will be discussed in Section 4.5 together with the results for performance.

### 4.4.3 Effects of Salience of Ability Grouping and Group Assignment on Performance

We will now analyze whether ability grouping affects participants' outcomes in the second test. First, we will compare the test score averages between people in the non-salient and the salient grouping treatment. Then we will look at the interaction effects between the assigned group and salience of group assignment on average test scores. Note that the bar graphs in Figure 4.6 are showing raw scores from the second test. As can be seen in Figure 4.6, there is neither a significant overall effect of salience of ability grouping nor an interaction effect of salience of ability grouping with group assignment on performance.

Figure 4.6: Effects of Salience of Ability Grouping and Group Assignment on Performance




*Note*: Panel A shows the effect of salience of ability grouping on test scores. Panel B shows the interaction effect of salience of ability grouping and group assignment on test scores.

In Table 4.3 the treatment effects of ability grouping on performance are analyzed separately for lower and higher ability subjects (who had below and above median performance, respectively, in the first test). Interestingly, we find opposite and significant effects for the two groups that are disguised when looking at the average over both groups as in Figure 4.6. As can be seen in Columns 1 and 2, while lower ability subjects perform significantly worse (by 3.1 points), higher ability subjects perform significantly better (by 2.7 points) with non-salient ability grouping than with salient ability grouping. Columns 3 and 4 show that when assigned to the weaker group, lower ability subjects benefit from learning that their group is the weaker one (by 4.1 points), while higher ability subjects suffer from learning that their group is the weaker one (by 6.6 points).[11]

Hence, we find that salient ability grouping has a positive effect on the performance of lower ability individuals while it has a negative effect on the performance of higher ability individuals. This is driven by opposite effects for these groups when

[11]We do not find that people's effort, in terms of revealing information by clicking on city names in the learning phase, which was meant to measure the intensity of their learning, responded to our treatments (see Table 4.7 in Appendix 4.7.3). We infer that subjects rather responded to the intervention by adjusting their mental efforts and that it may be better use of the revealed information that improves test outcomes.

they are assigned to the weaker group. While the performance of lower ability individuals increases when they learn that they were assigned to the weaker group, the performance of higher ability individuals decreases when they learn that they were assigned to the weaker group. This suggests that, in our setting, higher confidence in ability as measured by the learning test does not clearly result in better test outcomes. In fact, only for higher ability subjects confidence and subsequent performance are positively correlated, while they are negatively correlated for lower ability subjects. For the whole sample, confidence predicts subsequent outcomes negatively (p=0.027, see Table 4.8 in Appendix 4.7.3). Although we intended our feedback about performance in the "learning test" to influence subjects' beliefs about their marginal productivity of effort, which we expected to be positively related to effort, our feedback possibly (also) influenced a different type of belief.[12]

Table 4.3: Effects of Salience of Ability Grouping and Group Assignment on Performance

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable: Test Score | If Lower A. | If Higher A. | If Lower A. | If Higher A. |
| Non-salient Grouping | -3.053** | 2.743* | -4.067** | 6.581** |
| | (-2.14) | (1.74) | (-2.22) | (2.59) |
| Stronger Group | -1.783 | 0.406 | -2.984 | 4.080 |
| | (-1.13) | (0.26) | (-1.33) | (1.53) |
| Non-salient Grouping × Stronger Group | | | 2.528 | -6.293* |
| | | | (0.83) | (-1.82) |
| Observations | 76 | 81 | 76 | 81 |
| $R^2$ | 0.144 | 0.099 | 0.153 | 0.142 |

*Note*: This table presents the effect of non-salient versus salient ability grouping and assignment to a stronger versus a weaker group using a linear regression model including a constant, session fixed effects and robust standard errors. Dependent variable: test score. Columns 1 and 3 (2 and 4) show results for lower (higher) ability subjects. t statistics are reported in parentheses * p<0.10, ** p<0.05, *** p<0.01.

## 4.5 Discussion

We studied the causal effects of assignment to a weaker or a stronger group as well as its interaction with salience of the assignment mechanism on confidence in

[12]The belief we manipulated does not seem to be (only) a person's baseline belief in receiving a good outcome, which Fischer and Sliwka (Chapter 5 of this thesis) show may be negatively related to subsequent performance, because we find the inverse relationship for higher and lower ability subjects compared to what they find.

learning ability and outcomes in a subsequent learning test. To do so, we designed a feedback intervention that gave people imprecise feedback about either (1) their standing within their group (whether they performed in the upper or the lower half relative to their group) or (2) their standing within their group plus their group's standing relative to another group (whether their group was stronger or weaker than the other group). We expected, based on empirical research that finds that students become more confident in their academic abilities when they have worse classmates (Murphy and Weinhardt, 2014; Elsner and Isphording, 2017), that when only learning about their standing within their group, subjects would become more confident when they were assigned to the weaker group. Furthermore, when learning about both their standing within their group and their group's standing relative to another group, this effect should be expected to disappear if subjects assign correct weights to within-group and between-group ability signals, as in this case between-group information would counterbalance within-group information. However, if subjects overweigh between-group information, the effect of weaker group assignment would be negative, while if they overweigh within-group information the effect of weaker group assignment would still be positive.

Our results show that, in the setting we studied, when the group assignment mechanism was non-salient, it did not matter for subjects confidence whether they were assigned to the weaker or the stronger group. The signs of the effects suggest that in this case subjects were slightly more confident when assigned to the weaker group, however the effect sizes are so small that it would need a much larger sample size to possibly find a significant effect. Furthermore, we find that if the group assignment mechanism was salient, weaker group assignment made people less confident. This effect is highly significant and much larger than the positive effect of weaker group assignment when the assignment mechanism was non-salient. We find this effect both for lower and higher ability individuals, although it seems to be even stronger for the latter. We also find that subjects are on average less confident when the group assignment mechanism is salient than when it is non-salient.

This is found to be the case due to salient grouping causing subjects' beliefs to become decalibrated, especially when learning they were assigned to the weaker group. When grouping was non-salient, subjects on average gave quite correct estimates of their ability rank. However, when grouping was salient, subjects who were assigned to the stronger group were significantly overconfident while subjects who were assigned to the weaker group were significantly underconfident, indicating that people *overweighed* ability signals coming from *between-group* information.

When ability grouping was salient, subjects assigned to the weaker group were more underconfident than subjects assigned to the stronger group were overconfident, indicating that people *overweighed negative* information as compared to positive information. Some of those who are told they are in the weaker group report "irrational" rank beliefs (i.e. beliefs that must be false given the subject's information), while none of those who are told they were in the stronger group do so. When comparing people who received extreme feedback ("upper half in stronger group" and "lower half in weaker group") we find that, conditional on feedback, these groups did not have significantly different belief distributions although they represented the two extremes of the ability distribution. However, when it comes to ambivalent feedback, we find more decalibrated beliefs among those receiving bad between-group and good within-group information ("upper half in weaker group") than those receiving good between-group and bad within-group information ("lower half in stronger group"), which cannot be explained by lower abilities of the former group as compared to the latter. Thus, group assignment information seems to lead to stronger decalibration of beliefs if it is negatively surprising than if it is positively surprising. This is in line with the finding that people's beliefs respond more strongly to negative information (Ertac, 2011) but contradicts the possibly more common finding that people incorporate positive information into their beliefs more strongly than negative information (Eil and Rao, 2011; Mobius et al., 2011; Grossman and Owens, 2012; Wiswall and Zafar, 2015).

With respect to test outcomes, we find that salient ability grouping has a positive

effect on the performance of lower ability individuals while it has a negative effect on the performance of higher ability individuals. This is driven by opposite effects for these groups when they are saliently assigned to the weaker group. While the performance of lower ability individuals increases when they learn that they were assigned to the weaker group, the performance of higher ability individuals decreases when they learn that they were assigned to the weaker group. Past research has also variously found that performance increases (Kuhnen and Tymula, 2012; Azmat et al., 2016; Fischer and Wagner, Chapter 3 of this thesis) or decreases (Buser, 2016) in response to negative performance information. Our findings suggest that in our setting, higher confidence in learning ability as measured by the test does not have clear benefits for people in terms of improving their test outcomes. In fact, confidence overall predicts subsequent test outcomes negatively.

## 4.6 Conclusion

To our knowledge, this is the first study to investigate the causal effects of within-group and between-group information on people's ability beliefs and performance. Overall, our results suggest that ability grouping may have negative effects on people's confidence in their ability, especially for those who are assigned to a weaker group. Being part of a weaker peer group should not generally be expected to make people more confident. Our results imply that the positive effect of weaker peers on confidence if relative ability between groups is non-salient may be greatly outweighed by the negative effect of having weaker peers when people know that their peers are relatively weaker compared to another group. In line with past findings (Coffman, 2014), negative information about one's group may lead people to self-stereotype, i.e. to believe that one has worse characteristics than one actually does. Our results also suggest that, in settings where ability grouping is done visibly, forming ability groups may risk harming those people who are negatively surprised by weaker group assignment more than it may benefit those who are positively surprised by stronger group assignment.

The results of this study demonstrate that the effects of one's group's abilities on beliefs in own ability and subsequent performance are sensitive to information about the group assignment process. Because of this, one should be careful when interpreting effects of peer group ability on performance from field experiments where the group assignment mechanism is non-salient (as e.g. in Duflo et al., 2011; Carrell et al., 2013; Booij et al., 2017) as other effects may prevail once people find out that groups of different abilities were deliberately formed.

Overall, our findings suggest that the relationship between ability beliefs and motivation are complex and should be further investigated in future research. Our study may help to understand the effects of ability grouping in the field by isolating the effects it may have on ability beliefs. However, we caution that our results are based on a laboratory experiment that studies the effects in an abstract setting and further research needs to be done to confirm that our findings hold in educational or workplace settings.

## 4.7 Appendix to Chapter 4

### 4.7.1 Details on the Experimental Procedure

#### Test and Feedback Screens

Figure 4.7: Test 1 (Test Phase)

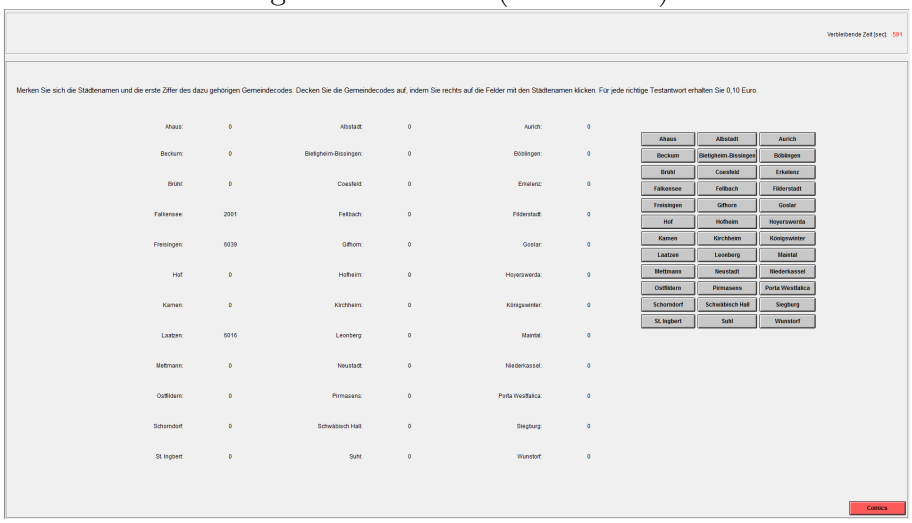



Figure 4.8: Test 1 (Learning Phase)

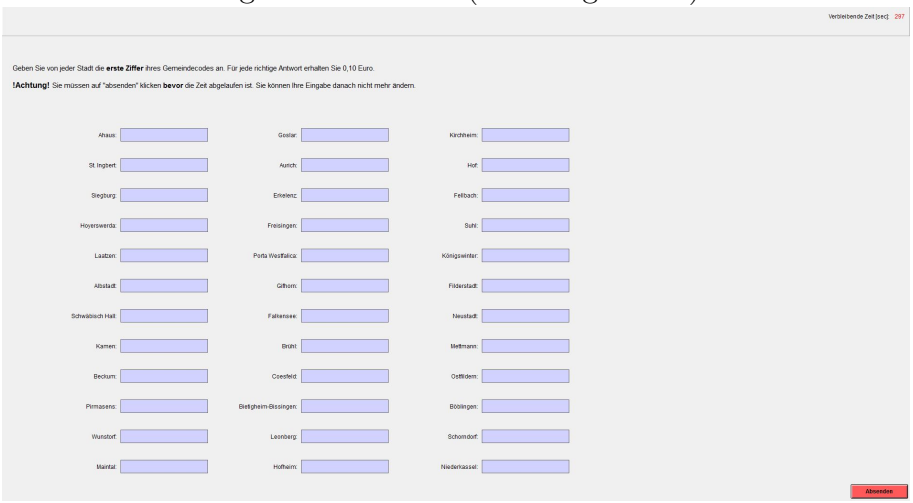


Figure 4.9: Sample Feedback: Non-salient Grouping

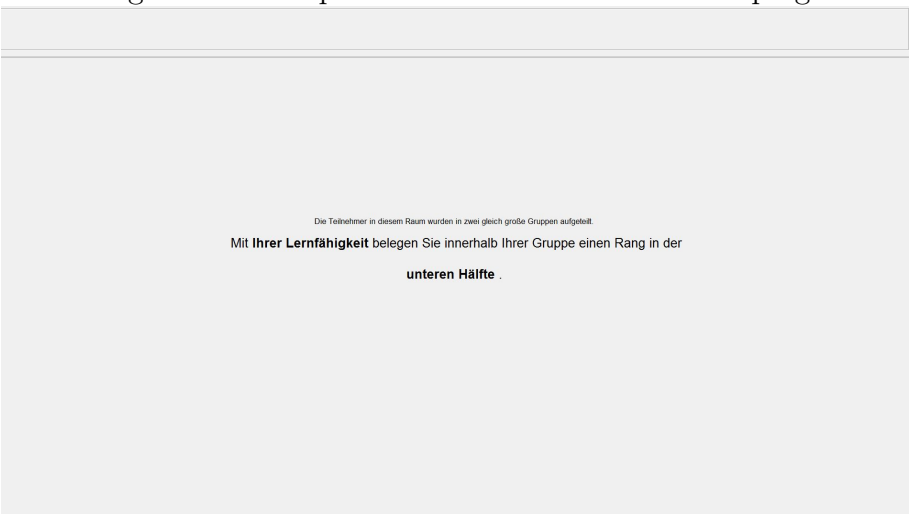



Figure 4.10: Sample Feedback: Salient Grouping

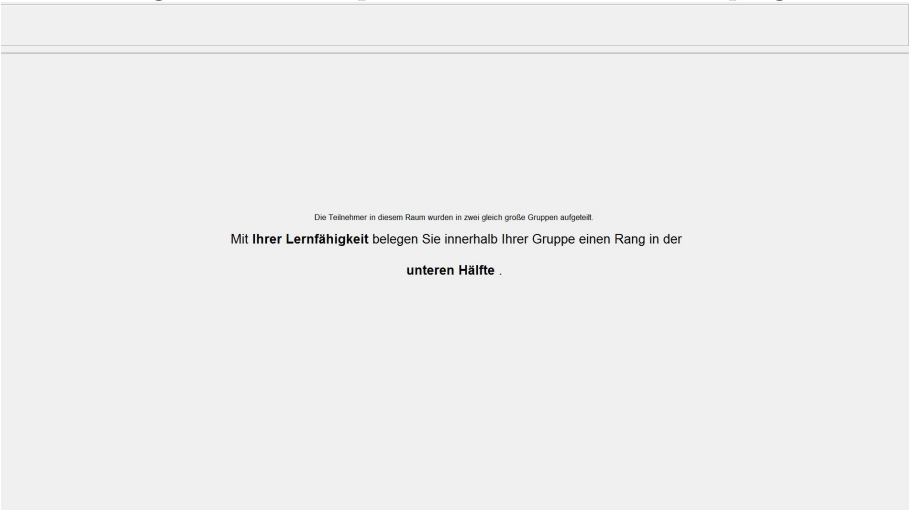



Figure 4.11: Sample Feedback: No Grouping

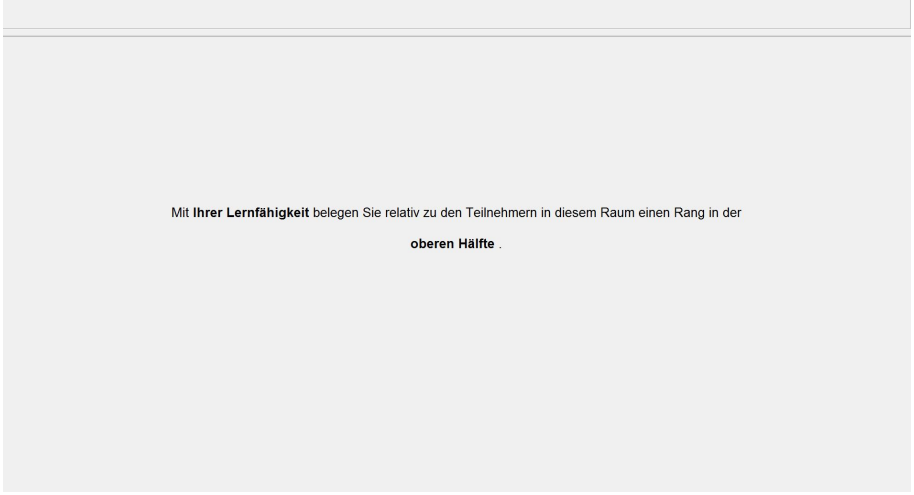


Figure 4.12: Test 2 (Test Phase)

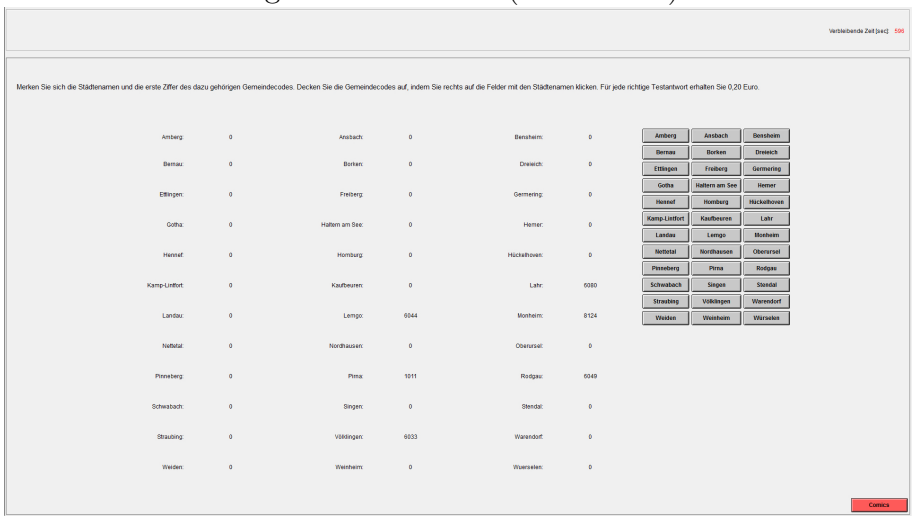
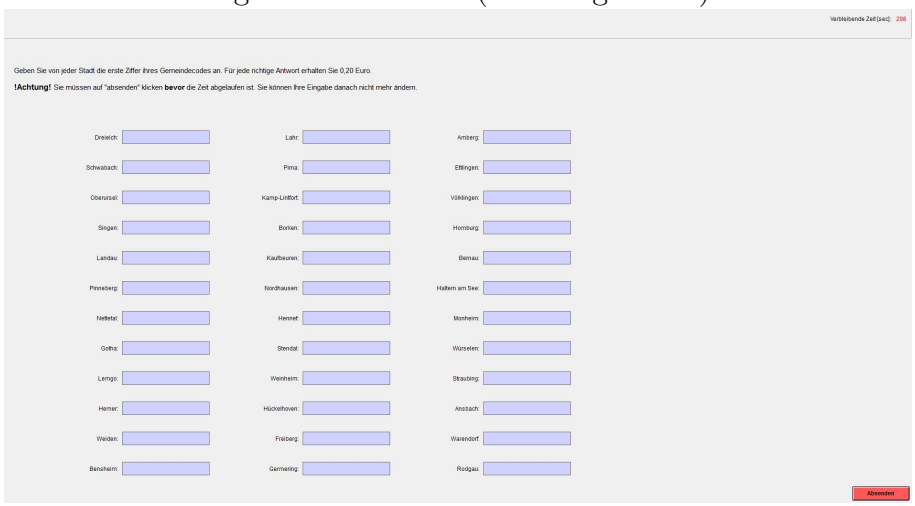



Figure 4.13: Test 2 (Learning Phase)



**Treatment Messages**

Table 4.4: Message by Treatment

| *Treatment:* | **Non-salient grouping** | **Salient grouping** |
|---|---|---|
| Message: | "The participants in this room were divided into two equally sized groups. With your learning ability you occupy a ranking in the upper [lower] half within your group." | "The participants in this room were divided into to equally sized groups: The stronger learners and the weaker learners. There, a better result makes it much more likely to be assigned to the stronger learners. You were assigned to the stronger (weaker) learners. With your learning ability you occupy a rank in the upper [lower] half among the stronger (weaker) learners." |

### 4.7.2 Summary Statistics and Balance Checks

Table 4.5: Summary Statistics

| Variable | Mean | Std. Dev. | Min. | Max. | N |
|---|---|---|---|---|---|
| Points Test 1 | 19.881 | 7.676 | 3 | 36 | 193 |
| Points Test 2 | 22.668 | 7.888 | 0 | 36 | 193 |
| Better Half | 0.508 | 0.501 | 0 | 1 | 193 |
| Confidence | -0.451 | 6.406 | -20 | 19 | 193 |
| Decalibration | 4.793 | 4.261 | 0 | 20 | 193 |
| Effort 1 | 239.539 | 117.775 | 51 | 898 | 193 |
| Effort 2 | 242.518 | 120.898 | 56 | 672 | 193 |
| Non-salient Tracking | 0.409 | 0.493 | 0 | 1 | 193 |
| Salient Tracking | 0.404 | 0.492 | 0 | 1 | 193 |
| Stronger Group | 0.497 | 0.502 | 0 | 1 | 157 |
| Better Half in Group | 0.409 | 0.493 | 0 | 1 | 193 |
| Extreme Feedback | 1.538 | 0.505 | 1 | 2 | 39 |
| Ambivalent Feedback | 1.462 | 0.505 | 1 | 2 | 39 |
| Female | 0.492 | 0.501 | 0 | 1 | 193 |
| Semester | 5.611 | 3.483 | 1 | 15 | 193 |
| School GPA | 2.574 | 6.424 | 0 | 90 | 193 |
| Profit | 11.41 | 2.295 | 5.8 | 17.6 | 193 |
| Session 1 | 0.145 | 0.353 | 0 | 1 | 193 |
| Session 2 | 0.135 | 0.342 | 0 | 1 | 193 |
| Session 3 | 0.155 | 0.363 | 0 | 1 | 193 |
| Session 4 | 0.135 | 0.342 | 0 | 1 | 193 |
| Session 5 | 0.15 | 0.358 | 0 | 1 | 193 |
| Session 6 | 0.119 | 0.325 | 0 | 1 | 193 |
| Session 7 | 0.161 | 0.368 | 0 | 1 | 193 |
| Humanities | 0.098 | 0.299 | 0 | 1 | 193 |
| Social Science | 0.078 | 0.268 | 0 | 1 | 193 |
| Law | 0.109 | 0.312 | 0 | 1 | 193 |
| Busines Administration | 0.295 | 0.457 | 0 | 1 | 193 |
| Economics | 0.161 | 0.368 | 0 | 1 | 193 |
| Medicine | 0.062 | 0.242 | 0 | 1 | 193 |
| Natural Sciences | 0.078 | 0.268 | 0 | 1 | 193 |
| Other Fields | 0.119 | 0.325 | 0 | 1 | 193 |

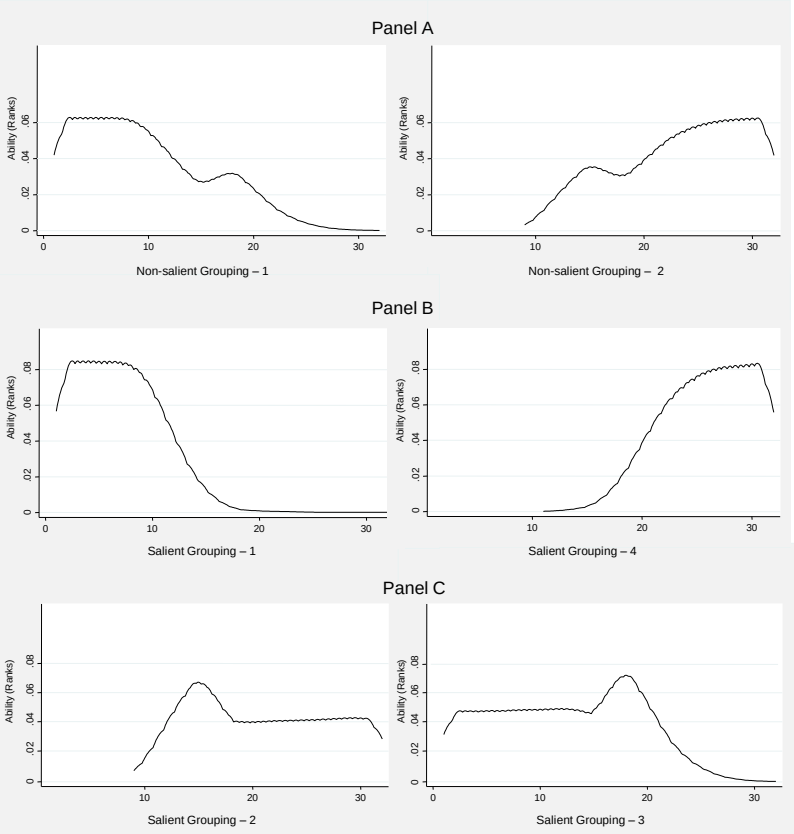
Table 4.6: Balance Check

| | (1) Non-Transparent Tracking | (2) Transparent Tracking | (3) No Tracking | (4) Overall | (1) vs. (2), p-value | (1) vs. (3), p-value | (2) vs. (3), p-value |
|---|---|---|---|---|---|---|---|
| Female | 0.481 | 0.513 | 0.472 | 0.492 | 0.692 | 0.931 | 0.690 |
| | (0.057) | (0.057) | (0.084) | (0.036) | | | |
| Points Test 1 | 20.418 | 19.628 | 19.250 | 19.881 | 0.513 | 0.478 | 0.801 |
| | (0.907) | (0.791) | (1.391) | (0.553) | | | |
| School GPA | 2.101 | 1.982 | 2.092 | 2.051 | 0.239 | 0.936 | 0.399 |
| | (0.068) | (0.075) | (0.100) | (0.045) | | | |
| Semester | 5.532 | 6.000 | 4.944 | 5.611 | 0.408 | 0.362 | 0.156 |
| | (0.358) | (0.438) | (0.534) | (0.251) | | | |
| Field of Study | 4.304 | 4.782 | 4.611 | 4.554 | 0.201 | 0.479 | 0.710 |
| | (0.252) | (0.274) | (0.322) | (0.163) | | | |
| Session No. | 4.177 | 3.872 | 3.944 | 4.010 | 0.357 | 0.566 | 0.856 |
| | (0.235) | (0.233) | (0.303) | (0.146) | | | |
| $N$ | 79 | 78 | 36 | 193 | | | |
| Proportion | 0.409 | 0.404 | 0.187 | 1.000 | | | |

Standard errors in parentheses.

### 4.7.3 Simulations and Further Results

Figure 4.14: Expected Ranks by Feedback Type




*Note*: This figure shows the distributions of the expected ability ranks by feedback type. The graphs are based on simulations applying our ability group assignment mechanism to 64,000 observations.

Table 4.7: Effort Intensity

| Dependent Variable: Effort | (1) If Lower Ability | (2) If Higher Ability | (3) If Lower Ability | (4) If Higher Ability |
|---|---|---|---|---|
| Non-salient Grouping | 29.90 | -13.18 | 34.34 | 17.08 |
| | (1.17) | (-0.46) | (0.97) | (0.38) |
| Stronger Group | -43.74 | -14.50 | -38.48 | 14.46 |
| | (-1.56) | (-0.52) | (-1.06) | (0.32) |
| Non-salient Gr. × Stronger Group | | | -11.06 | -49.62 |
| | | | (-0.20) | (-0.76) |
| Observations | 76 | 81 | 76 | 81 |
| $R^2$ | 0.181 | 0.128 | 0.181 | 0.135 |

*Note*: This table presents the effect of non-salient versus salient ability grouping and assignment to a stronger versus a weaker group using a linear regression model including a constant, session fixed effects and robust standard errors. Dependent variable: effort in terms of clicks. Columns 1 and 3 (2 and 4) show results for lower (higher) ability subjects. t statistics are reported in parentheses * p<0.10, ** p<0.05, *** p<0.01.

Table 4.8: Correlation between Confidence and Subsequent Performance

| Dependent Variable: Test Score | (1) All | (2) If Lower Ability | (3) If Higher Ability |
|---|---|---|---|
| Confidence | -0.205** | -0.148 | 0.0312 |
| | (-2.32) | (-1.27) | (0.24) |
| Observations | 157 | 76 | 81 |
| $R^2$ | 0.081 | 0.096 | 0.064 |

*Note*: This table presents the correlation between confidence and subsequent performance using a linear regression model including a constant, session fixed effects and robust standard errors. Dependent variable: test score. Column 1 shows results for all subjects, and columns 2 and 3 show results for lower and higher ability subjects, respectively. t statistics are reported in parentheses * p<0.10, ** p<0.05, *** p<0.01.

# Chapter 5

# Confidence in Knowledge or Confidence in the Ability to Learn: An Experiment on the Causal Effects of Beliefs on Motivation

Co-authored with Dirk Sliwka[1]

## 5.1 Introduction

Motivational beliefs are held to be a strong determinant of important life outcomes such as educational attainment and professional development. However, there seems to be disagreement in the public realm on which beliefs about ourselves are beneficial for us. Folk wisdom tells us that holding a very favorable opinion of our abilities may often breed failure as it tempts us to rest on our laurels and lowers our motivation to work hard towards our goals and the economics literature, too, mostly emphasizes the negative effects of too much confidence. However, many popular self-help books claim that increasing our self-confidence makes us more likely to be successful in life.[2] In educational settings, optimistic beliefs about ourselves are widely thought to foster skill development and a quick search on the internet will turn up many school-related websites and workshop offers claiming that fostering children's confidence will improve their motivation to learn. However, there seems to be disagreement about whether praise for performance, effort, or progress is best

[1]My co-author and I contributed equally to the design and implementation of the study, to the data analysis, and to writing the paper. The theoretical model is by my co-author.

[2]The claim "confidence breeds success" produces 329 hits on Google Books and a search on Amazon.com for "confidence" in the sub-category "Books - Self-Help - Success" produces 783 hits.

to raise confidence and motivation to learn.

A straightforward conjecture is that some of the disagreement in the popular discourse about the relationship between feedback, confidence, and performance is caused by the tendency to subsume different types of beliefs under the notion of "confidence". Different types of feedback may influence beliefs about different dimensions of a person's skills and abilities and conditional on circumstances a shift in a belief about a given skill dimension may or may not raise motivation to exert effort.[3]

The key purpose of this paper is to distinguish two dimensions of confidence – *confidence in one's level of prior knowledge* and *confidence in one's learning ability* – and to study causal effects of changes in these dimensions of a person's confidence on investments in human capital. Reinforcement of confidence in these two dimensions likely has very different effects, as the first dimension is related to one's ex-ante probability of passing a test while the second one is related to how much one's passing probability increases when exerting learning efforts. We first illustrate these belief dimensions in a simple formal model and then study the effects of exogenous variation in both dimensions in a lab experiment.

The motivational role of confidence has attracted substantial interest from different fields in economics in recent years. Bénabou and Tirole (2002, 2003), for instance, have studied formal models in which agents are uncertain about the marginal returns to their effort. These models yield a precise notion of confidence as an agent's belief in her own marginal product of effort. A higher confidence then naturally induces an agent to work harder on a task.[4] The recent literature on the economics of education has studied specific personality traits that predict important life outcomes (Heckman et al., 2006; Cebi, 2007; Heineck and Anger, 2010; Heckman and Kautz, 2012). Internal locus of control and self-esteem, psychological constructs intended to

[3]Indeed, the literature in psychology indicates that there is mixed evidence on the association between different types of feedback and performance (Kluger and DeNisi, 1998; Hattie and Timperley, 2007).

[4]See, for instance, Koch et al. (2015) for an overview on these and related models from the perspective of the economics of education.

capture a person's beliefs about the ability to affect outcomes, feature prominently among these traits. There is also empirical evidence that socially disadvantaged children (Filippin and Paccagnella, 2012), and girls (Reuben et al., 2017) are less confident about their academic ability and that this has negative effects on their educational decisions and expected earnings.

In our experiment students have to decide how intensively they want to prepare for a test. They pass the test and earn a reward if their performance reaches a certain threshold. Based on the analysis of a simple formal model we hypothesize that a higher *confidence in the level of prior knowledge* causes students with low levels of knowledge to invest more. This is because it subjectively moves them closer to the passing threshold and raises the probability that an additional remembered item is pivotal to passing the test. For students with high levels of prior knowledge we expect the opposite, i.e. that raising their confidence in knowledge even further will lower their effort to prepare for the test because it subjectively moves them further away from the passing threshold such that learning becomes less relevant for whether someone passes or fails the test. For the other dimension – *confidence in learning ability* – we expect that raising this dimension of confidence will have a monotonic effect and cause students to invest more effort in learning because the perceived marginal cost of effort to generate "knowledge" decreases.

To study the causal effects of the two dimensions of confidence, we exogenously vary feedback scores subjects receive about their performance in two prior tests. One of these tests measures their prior knowledge, the other test measures the ability to memorize information. After completing these two tests, each subject privately receives a feedback score for each of the tests. Subjects know that each feedback score is the sum of their true score in the respective test and a random noise term. We then elicit subjects' confidence by asking them to estimate their own rank in the first two tests. Subjects can then buy pieces of information and memorize these to prepare for a final test in which they earn a fixed amount of money if their performance exceeds a specific threshold. The random component

in the feedback scores thus generates exogenous variation in the agents' confidence in the two dimensions, which we use as instrumental variables to estimate causal effects of confidence on investment decisions and test outcomes.

We find that a higher confidence in learning raises learning investments irrespective of the prior level of knowledge. Confidence in knowledge, however, has a negative effect on investments of individuals with above average prior knowledge and a positive effect on investments of individuals with below average prior knowledge. With respect to test outcomes, we find that raising the confidence in learning of individuals with below average prior knowledge improves their rank in the final test and their probability of passing it, however, we do not find a beneficial effect for individuals who already had above average prior knowledge. Mirroring the effects of confidence in knowledge on effort, we find that raising confidence in knowledge of individuals with above average prior knowledge decreases their outcomes in the final test whereas it has the opposite effect on individuals with below average prior knowledge.

This paper makes two contributions. First, it shows theoretically and experimentally that in situations where choices involve effort, confidence should be viewed as a multidimensional concept (even if the effort choice is unidimensional) and that general statements about the motivational effects of confidence are misleading. In order to explain the effects of confidence on motivation to exert effort, and on learning in particular, we have to understand which roles effort and ability play in achieving a goal. An important implication of this is also that interventions aimed at raising confidence should be carefully designed and evaluated because they might affect several beliefs that interact in different ways with motivation to exert effort. Second, we develop a deception-free experimental approach to study the *causal* effect of beliefs on effort by generating exogenous variation in two dimensions of confidence. For this reason, we can rule out that, for instance, unobserved psychological dispositions that may be correlated with confidence drive the association between confidence, motivation to exert effort, and performance. By studying the effects of confidence on

learning decisions and test outcomes, our study links the literature on experiments in education to the literature on motivational beliefs and socio-emotional skills.

The remainder of the paper is structured as follows. Section 2 summarizes the related literature on the determinants of effort provision in educational and similar settings. Section 3 presents a model and derives best responses and hypotheses from it. Section 4 presents the experimental design. Section 5 presents the results and Section 6 concludes.

## 5.2 Related Literature

Our research is closely related to the game theoretical and behavioral economic literature on confidence and incentives. As stated above, "confidence in learning ability" in our setting is equivalent to Benabou and Tirole's (2002, 2003) notion of confidence as an agent's (rational) belief in her own marginal product of effort. We study the interplay between this type of confidence and confidence in prior knowledge as well as the impact of both on investment incentives.[5]

The effects of beliefs in and feedback about ability have been explored in several theoretical papers. The role of feedback in tournament settings has, for instance, been explored by Aoyagi (2010) and Gershkov and Perry (2009). Most closely related to our study is the analysis by Ederer (2010) who studies the effect of interim feedback (about interim outcomes) on effort and shows that when effort and ability are complements feedback should induce competing effects as it informs agents about their relative standing (which reduces incentives) as well as their ability (which may increase incentives). In a principal-agent setting, Santos-Pinto (2008) shows that a worker's overestimation of his ability is beneficial for the principal when ability and effort are complements but not when they are substitutes. Our experiment provides causal empirical evidence for the relevance of disentangling different ability beliefs.

In the context of job search on the labor market, contributions by Caliendo

[5] Compte and Postlewaite (2004) depart even further from a neoclassical framework by assuming that confidence, influenced by an agent's past successes and failures, raises the (factual) probability of success of an agent.

et al. (2015) and Spinnewijn (2015) have studied the role of different dimensions of confidence on search efforts. Most closely related to our model is the analysis of Spinnewijn (2015), who studies how biased beliefs in two dimensions influence job search: "baseline beliefs" – the beliefs about the baseline job finding probability for given search efforts, and "control beliefs" – the beliefs about the increase in the job finding probability when searching more intensively. We study the effect of baseline belief (concerning prior knowledge) and control belief (concerning ability to learn) on learning effort and provide causal evidence on their impact in an educational setting.

A number of empirical and experimental papers have studied the effect of feedback about (relative) performance on educational outcomes. Tran and Zeckhauser (2012) find that students perform significantly better in a final English test when they are told their rankings on practice tests than students in the control group who only receive private feedback about their test score. Bandiera et al. (2015) exploit rule differences between university departments concerning the provision of feedback to students and find that students who receive their individual exam grade prior to writing a long essay do better in it than students who do not. Azmat and Iriberri (2010), in a natural field experiment set in a high school, find that students who repeatedly receive information about the average grade of their class in addition to information about their own grade, receive 5 percent better grades. In Azmat et al. (2016), however, a random sample of college students who receive information about their position in the distribution of grades repeatedly over a period of three years are found to do worse during the first six months. As the authors argue, students in their sample were initially underconfident. Thus learning that they were doing better than expected had a negative impact on performance. In line with this argument, Kuhnen and Tymula (2012), who study effort reactions to rank feedback in the lab, find that individuals who ranked better than expected decrease output, whereas those who ranked worse than expected increase output. In contrast to these studies, we do not vary feedback on the relative rank in the relevant test but go one

step back and manipulate the beliefs a person holds about her knowledge and ability to learn in order to shed light on the behavioral mechanisms by which feedback affects behavior.

Finally, although incentive compatible measurement of beliefs is common in economic laboratory studies, there are very few studies which generate *exogenous* variation in beliefs in order to study the causal effect of beliefs on actions. Mobius et al. (2011) repeatedly give noisy feedback about whether one performed in the better or the worse half of participants in an IQ test. The authors use the random variation in the feedback to estimate the causal effect of confidence in own ability on the aversion to receiving information about ability and find that a lower confidence induces a stronger aversion to receiving information about one's own ability. Schwardmann and Van der Weele (2016) investigate the hypothesis that overconfidence serves to more effectively persuade others and also manipulate subjects' confidence in their own intelligence using noisy feedback. Costa-Gomes et al. (2014) study the causal effect of beliefs in a trust game by inducing a zero-mean random shift that exogenously increases or reduces the trustee's level of re-payment. Then the authors use the random shift as instrumental variable to estimate the causal effect of beliefs about the trustee's transfer share on the trustor's choice. Our study is the first that uses noisy feedback to manipulate two different belief dimensions in order to study the causal effect of ability beliefs on learning investments and test outcomes.

## 5.3 An Illustrative Model

Consider the following simple illustrative model which can be interpreted as an analysis of a reaction function in a standard Lazear and Rosen (1981) tournament in which we allow the agent's beliefs to vary with respect to (i) the costs of effort (ability $a$) and (ii) a potential handicap/or lead (prior knowledge $k$). In contrast to the standard tournament literature we do not analyze the equilibrium behavior of a small set of players but follow Casas-Arce and Martínez-Jerez (2009) in studying a "population tournament" where the threshold necessary to win the prize is deter-

ministic. The model's purpose is to illustrate how changes in these two forms of "confidence" should affect the efforts exerted to win the prize.

A risk neutral agent can invest effort to raise her human capital. Human capital is measured by "pieces of knowledge". An agent's posterior knowledge is the sum of her prior knowledge $k$ and knowledge acquired through learning $\Delta$. Knowledge acquisition is costly and the agent's cost function is

$$c(\Delta, a)$$

where $a$ measures the agent's ability to acquire further knowledge. We assume that $\frac{\partial c}{\partial \Delta}, \frac{\partial^2 c}{\partial \Delta^2} > 0$ and $\frac{\partial c}{\partial \Delta \partial a} < 0$ such that the marginal costs of knowledge acquisition are smaller for more able agents. The agent is uncertain about both, her prior knowledge $k$ and the ability to acquire further knowledge $a$. She knows that both are distributed according to the cumulative distribution functions $F_a(a)$ and $F_k(k)$. The agent receives informative signals $s = (s_a, s_k)$ such that $\frac{\partial E[a|s_a,s_k]}{\partial s_a} > 0$ and $\frac{\partial E[k|s_a,s_k]}{\partial s_k} > 0$. Note that we can decompose

$$\begin{aligned} a &= E[a|s_a, s_k] + \varepsilon_{as} \\ k &= E[k|s_a, s_k] + \varepsilon_{ks} \end{aligned}$$

where $\varepsilon_{as}$ and $\varepsilon_{ks}$ are uncorrelated with the signals $(s_a, s_k)$ and have mean zero (by the law of iterated expectations).[6] Assume that $\varepsilon_{as}$ and $\varepsilon_{ks}$ have unimodal densities with $g'_{\varepsilon_{as}}(0) = g'_{\varepsilon_{ks}}(0) = 0$. Denote the conditional expectations as

$$\begin{aligned} \hat{k} &= E[k|s_a, s_k] \\ \hat{a} &= E[a|s_a, s_k] \end{aligned}$$

such that $\hat{k}$ and $\hat{a}$ describe the agent's own mean belief in her knowledge and costs of

[6] To see, for instance, that $Cov[s_a, \varepsilon_{as}] = Cov[s_a, a - E[a|s_a, s_k]] = 0$ note that by the law of iterated expectations $E[s_a(a - E[a|s_a, s_k])] = E[E[s_a(a - E[a|s_a, s_k])|s_\gamma, s_k]] = E[s_a E[(a - E[a|s_a, s_k])|s_a, s_k]] = 0$.

knowledge acquisition respectively. The decomposition allows us to do comparative statics with respect to $\hat{k}$ and $\hat{a}$, which capture an agent's confidence in the two dimensions.

The agent attains a certain educational outcome, such as passing an admission test to an education program, or being awarded an academic title, if $k + \Delta$ exceeds a threshold value $\tau$.[7] In this case she will receive a reward $B$. The agent's objective function can thus be denoted as

$$\max_{\Delta} \Pr\left(\hat{k} + \varepsilon_{ks} + \Delta > \tau\right) B - E\left[c\left(\Delta, a\right) | s_a, s_k\right].$$

In order to guarantee that this optimization problem has a unique solution we assume that

$$\max_{\varepsilon}\left(-g'_{\varepsilon_{ks}}\left(\varepsilon\right)\right) B < \min_{\Delta,a} E\left[\frac{\partial^2 c\left(\Delta, a\right)}{\partial \Delta^2}\right] \tag{5.1}$$

which will, for instance, hold if $\frac{\partial^2 c(\Delta,a)}{\partial \Delta^2}$ is bounded from below by a constant and the signal $s_k$ is not too precise.[8]

The first derivative of the objective function is

$$g_{\varepsilon_{ks}}\left(\tau - \hat{k} - \Delta\right) B - E\left[\frac{\partial c\left(\Delta, \hat{a} + \varepsilon_a\right)}{\partial \Delta}\right]$$

and by condition (5.1) the objective function is strictly concave. We can now show:

**Proposition 1** *Knowledge acquired through learning $\Delta\left(\hat{a}, \hat{k}\right)$ is strictly increasing in the agent's confidence in her ability to acquire knowledge $\hat{a}$. It is strictly increasing in the agent's confidence in prior knowledge $\hat{k}$ if and only if $\hat{k}$ is smaller than a cut-off value and otherwise strictly decreasing.*

[7] Note that here we treat $\tau$ as an exogenous constant. If we consider a tournament setting $\tau$ will be determined in equilibrium by the choices of the other agents. In a tournament between a continuum of agents where a fixed fraction can win a prize the equilibrium threshold will indeed be deterministic (see, for instance Casas-Arce and Martínez-Jerez (2009)).

[8] This condition will guarantee that the objective function is strictly concave. Intuitively, if there is sufficient uncertainty on $k$ then $\varepsilon_{ks}$ will have a large variance. If, for instance, $\varepsilon_{ks}$ is normally distributed a large enough variance will guarantee that the slope of the density function will not be too steep.

**Proof:**

By implicit differentiation we obtain

$$\frac{\partial \Delta\left(\hat{a},\hat{k}\right)}{\partial a} = -\frac{-E\left[\frac{\partial c(\Delta,\hat{a}+\varepsilon_a)}{\partial \Delta \partial a}\right]}{-g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B - E\left[\frac{\partial^2 c(\Delta,\hat{a}+\varepsilon_a)}{\partial \Delta^2}\right]} > 0$$

as the denominator is negative by condition (5.1). And

$$\frac{\partial \Delta\left(\hat{a},\hat{k}\right)}{\partial \hat{k}} = -\frac{-g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B}{-g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B - E\left[\frac{\partial^2 c(\Delta,\hat{a}+\varepsilon_a)}{\partial \Delta^2}\right]} \tag{5.2}$$

such that

$$\frac{\partial \Delta\left(\hat{a},\hat{k}\right)}{\partial \hat{k}} > 0 \Leftrightarrow g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right) < 0$$

which, as $g_{\varepsilon_{ks}}(\varepsilon)$ has a unique mode at 0, is equivalent to

$$\tau > \hat{k} + \Delta\left(\hat{a},\hat{k}\right).$$

The right hand side is strictly increasing $k$ as $\frac{\partial \Delta\left(\hat{a},\hat{k}\right)}{\partial k} > -1$. To see the latter, note that

$$\frac{\partial \Delta\left(\hat{a},\hat{k}\right)}{\partial \hat{k}} = -\frac{-g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B}{-g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B - E\left[\frac{\partial^2 c(\Delta,\hat{a}+\varepsilon_a)}{\partial \Delta^2}\right]} > -1 \Leftrightarrow$$

$$g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B < g'_{\varepsilon_{ks}}\left(\tau-\hat{k}-\Delta\right)B + E\left[\frac{\partial^2 c\left(\Delta,\hat{a}+\varepsilon_a\right)}{\partial \Delta^2}\right]$$

which always holds. Hence, condition (5.2) holds for sufficiently small $k$ and will not hold above a threshold level.[9] ■

To illustrate the result, consider the following parametric example. Assume that the agent's cost function is $c(\Delta,a) = \frac{c-a}{2}\Delta^2$ and that the agent believes that $k$ is normally distributed with mean $\hat{k}$ and variance $V[\varepsilon_{ks}] = \sigma^2_{\varepsilon_k}$. As the cost function is linear in $a$, expected costs are equal to $\frac{c-\hat{a}}{2}\Delta^2$. The agent's objective function is

[9]Note that this threshold will be strictly positive if $\tau > k + \Delta\left(\hat{a},\hat{k}\right)$ for $k = 0$. A sufficient condition for this is that the objective function is downward sloping in $\Delta$ at $\Delta = \tau$ for $k = 0$, which is the case when $g_{\varepsilon_{ks}}(0)\,B < E\left[\frac{\partial c(\Delta,\hat{a}+\varepsilon_a)}{\partial \Delta}\right]$. This will hold if the signal on $k$ is not too precise.

thus

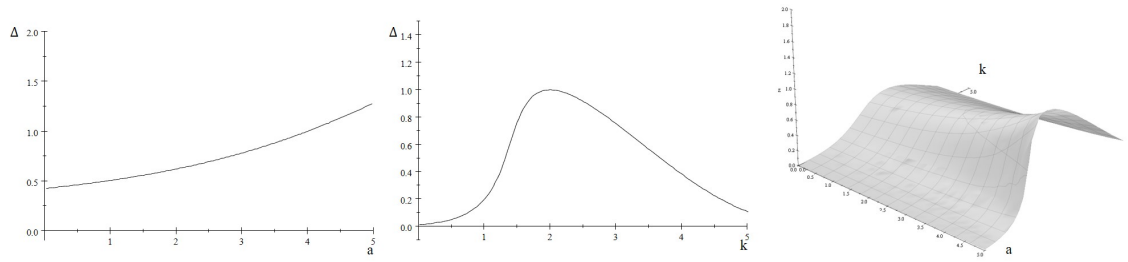
$$\max_{\Delta} \Pr\left(\varepsilon_{ks} > \tau - \Delta - \hat{k}\right) B - \frac{c - \hat{a}}{2}\Delta^2.$$

The first derivative of the objective function[10] becomes

$$\frac{1}{\sigma_{\varepsilon_k}}\phi\left(\frac{\tau - \Delta - \hat{k}}{\sigma_{\varepsilon_k}}\right) B - (c - \hat{a})\,\Delta = 0,$$

where $\phi(\varepsilon)$ is the pdf of a standard normal distribution. While this equation has no closed form solution we can use this expression to plot $\Delta$ as an implicit function of $a$ and $\Delta$ for specific examples.[11]

Figure 5.1: Learning Investments as a Function of Perceived Ability and Knowledge




Hence, *a higher confidence in the ability to learn always leads to higher learning investments* as it lowers the perceived marginal costs of learning efforts. This is essentially the motivational effect of self-confidence stressed, for instance, by Benabou and Tirole (2002). However, *confidence in prior knowledge* has a positive effect only for agents with low prior knowledge but reduces the incentives to learn for those with higher prior knowledge. The intuition is the following: If an agent has rather low confidence in her initial knowledge she thinks that the likelihood of achieving the educational outcome is small. In turn, the expected marginal gains from learning are small. Raising the confidence in knowledge raises the perceived likelihood to jump the threshold and consequently increases the marginal returns to learning efforts.

[10] Condition (5.1) that guarantees an internal solution here becomes $\frac{1}{\sigma^2_{\varepsilon_k}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}B < c - \hat{a}$, i.e. the objective function will be strictly concave if $\hat{a}$ is not too large.

[11] The plots use values $B = 10$, $\sigma^2_{\varepsilon_k} = 1$,
$\tau = 3$, $c = 8$ and the condition guaranteeing a strictly concave objective function requires that $\hat{a} < 8 - \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}10 = 5.5803$.

If, however, the agent believes that she has a very high level of prior knowledge, her perceived likelihood of attaining the outcome even at lower learning investments increases. In turn, the incentive to invest in acquiring further knowledge decreases.

Based on this illustrative model, we designed an experiment that enables us to clearly disentangle confidence in prior knowledge and confidence in the ability to learn and allows us to measure the causal effect of confidence in both dimensions.

## 5.4 Experimental Design

We have to keep in mind that confidence is inherently an endogenous variable as it will always be affected by unobserved experiences, abilities, and other traits of the respective subjects, which could also affect the outcome variables through different unobserved behavioral channels. Hence, merely detecting a correlation between confidence and behavior does not allow to infer causality. In order to avoid this problem, we have developed an experimental design in which we generate *instrumental variable*s, that is variables that are (i) cleanly exogenous but (ii) directly affect confidence. We then use these variables to investigate the causal effects of confidence on behavior. In the following we will explain in detail how we implemented this idea.

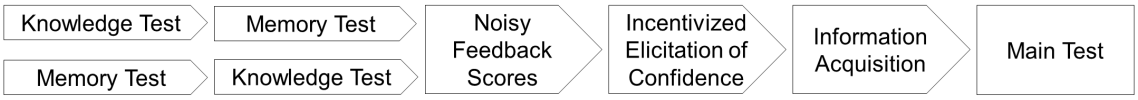
We invited university students to the Cologne Laboratory for Economic Research.[12] Upon arrival, registered participants were randomly assigned a computer. Before the experiment started, students were informed that they were prohibited to talk to each other, to use electronic devices or pen and paper during the experiment and that anyone who violated this rule would be excluded from the experiment. We monitored compliance with the rule during the entire session. Participants were informed that they would receive the regular show-up fee of 2.50 euros and that they could earn additional money during the experiment.[13]

[12]The laboratory uses the recruitment software ORSEE (Greiner, 2004) for managing the subject pool. The experiment was programmed using z-Tree (Fischbacher, 2007). Financial support of the Deutsche Forschungsgemeinschaft (DFG) through grant FOR1371 is gratefully acknowledged.

[13]A detailed description of the experiment's timeline, tests, feedback, and belief elicitation can be found in appendices 5.7.5 and 5.7.6.

The timeline of the experiment is illustrated in Figure 5.2 and can be summarized as follows: Before the main intervention, subjects take part in a memory and a knowledge test. Then they learn a feedback score about their performance in each test and these feedback scores are the sum of the respective test outcomes and random noise terms. Hence, this stage constitutes our treatment variation: The noise terms exogenously vary information that should affect subjects' confidence in the two dimensions. In a next step we elicit subjects' beliefs about their relative standing in both domains, which was incentivized by paying them for accuracy of beliefs. These are the main belief variables we use in our analysis as measures of confidence in the two dimensions. Then subjects can undertake a costly investment in further knowledge to prepare for a final test in which they can earn a substantial amount of money when passing a threshold. The learning investment as well as the test results will constitute our outcome variables.

Figure 5.2: Timeline of the Experimental Procedure




### 5.4.1 Stages of the Experiment

Measurement of Prior Knowledge and Learning Ability: After the introduction, participants saw a description of the test they were about to take first, which was either a *"knowledge test"* or a *"memory test"*. The order of the tests was randomized within each session to eliminate possible order effects. In the *knowledge test* subjects had to rank 60 cities according to their numbers of inhabitants within triples of cities, i.e. they had to state which city is the largest and which one is the smallest among three cities and would earn a piece rate of 0.10 euros for each correct set. In the *memory test* subjects first saw a list of 36 cities with a (fictitious) city code belonging to each city. This list was displayed on the screens for 15 minutes and subjects were not allowed to take notes. After this they had to rank cities within triples according to these city codes and would earn 0.20 euros for each correct set.

Hence, the knowledge test measured subjects' prior knowledge and the memory test measured their capacity to memorize information. The memory test closely resembles tests used by psychologists to test working memory capacity (Wilhelm et al., 2013) and was designed such that it covers the same domain (numbers attached to city names) as the knowledge test and in order to make one's performance in it seem as relevant as possible with respect to one's later learning decision for a test in this domain.[14]

Both tests were incentivized with a piece rate. Participants took the two tests one after another and after each test were asked how many triples they believed to have solved correctly, immediately afterwards they were also asked how many triples they believed other participants on average solved correctly. In both cases answers were not incentivized and participants were informed that their answer did not have any effect on the further course of the experiment. A detailed overview of the tests and stages of the experiment can be found in Appendices 5.7.5 and 5.7.6.[15] Then participants were informed there will be a "Test 3 (main test)", and that, unlike in the first two tests, they would earn 10 euros if they performed better than half of participants in the session who did the tests in the same order as them. They were also informed that they could prepare for this third test.

Feedback stage: Participants were informed that before preparing for the third test, they would receive feedback about their outcomes in the first two tests in the form of a *"knowledge score"* and a *"memory score"*. As explained to the participants, each score was the sum of a participant's number of correct sets in the respective test and a noise term uniformly and independently distributed between

[14]Working memory capacity is a strong predictor of ability to acquire knowledge and new skills, independently of IQ (Alloway and Alloway, 2010). See Ackerman et al. (2005) for an overview.

[15]We measured beliefs twice. Once before giving feedback (unincentivized) and once afterwards (incentivized - see details below). Note that the beliefs elicited after the feedback intervention are crucial for our design, as they serve as a measure of confidence that can be affected by the treatment intervention (i.e. the noisy feedback). We use the unincentivized measures only descriptively in order to evaluate how certain subjects were about their test outcomes. In fact, the correlation between beliefs about performance in the memory test (elicited after the test but before the feedback score was give) and in the knowledge test are 0.53 and 0.18, respectively. Thus, uncertainty about own ability and knowledge is generally high prior to learning the test score – which is a precondition for our feedback manipulation to work.

-2 and +2 such that each of the values (-2, -1, 0, 1, 2) is drawn with a probability of 20 percent and added to the true score.[16] The *randomly distributed noise term* thus creates exogenous variation in feedback about knowledge and learning ability while avoiding any form of deception. Then the personal feedback scores and average feedback scores of participants in past sessions were displayed on the same screen.[17] As already noted above, the exogenous variation in the personal feedback scores allows for the estimation of causal effects of the agents' confidence on behavior, a central contribution of our study, and thus an important design feature of our experiment.

Measurement of confidence: Participants were asked to estimate their rank in the knowledge and in the memory test relative to those participants in the room who worked on the two tests in the same order as them. They were informed that they could earn one euro, respectively, for estimating their rank in each test correctly.[18] Our design thus allows us to measure both the perceived level of ability (which is the focus of many economic studies of situations where a choice does not entail a decision about effort), and the perceived effectiveness of effort to raise the level of ability (the focus mainly of psychological studies employing non-incentivized questionnaires to measure self-efficacy and locus of control (Eccles and Wigfield, 2002)) in an incentive compatible manner.

Investment stage: After participants learned their knowledge score and their memory score they were shown a screen explaining the main *"combined knowledge and memory test"* in detail. Participants were informed that this test was based on

[16]For a similar approach compare, for instance, Grossman and Owens (2012) who study agent's reactions to noisy feedback about their own performance. Note that the incentives in rank order tournaments are not affected by random noise (For a summary of the literature see Dechenaux et al. (2015).)

[17]We always displayed the same average results from a pilot study to keep the frame of reference of the personal feedback constant between the experimental sessions. Participants in the pilot study were recruited from the same subject pool as participants in the experiment and results were very similar.

[18]This method is easy to explain and elicits the mode of an agent's subjective beliefs in an incentive compatible manner and is robust to risk aversion. To see that, note that an agent who has to state an estimate $r$, the value of a random variable $x$, and receives 1 euro when reporting correctly should report $argmax_r Pr\left(r = x\right) u(1) + \left(1 - Pr\left(r = x\right)\right) u(0)$, which is equal to the mode of the distribution. Since the range of beliefs in our context is small due to a limited number of ranks, the chances of having an exact estimate are reasonable.

the same field of knowledge and had the same length and structure as the initial knowledge test, i.e. they would have to rank sets of three cities according to the size of their populations. This time, however, they would earn a prize of 10 euros when doing better in this test than half of participants in the session who did the first two tests in the same order as them. Furthermore, they were told that they could prepare for it by acquiring information relevant to pass the test. To be specific, subjects had a budget of 3 euros to buy information about cities' numbers of inhabitants in packages of 10 cities for 0.5 euros per package. They could buy a maximum of 6 packages, together covering all the cities in the test. The decision on how many packages to buy was a one-shot decision, i.e. subjects had to state in advance how many packages they wanted to acquire [19] They knew that all cities they could "buy" were part of the later test and each package – when fully memorized – would allow to completely answer at least 3 assignments (triples) in the later test. The acquired packages were then displayed in a 15 minutes learning phase before the final test. In this phase subjects also had the possibility to click on a button in order to look at cartoons displayed on the screen (and subjects knew this before they acquired information).[20] Hence, subjects faced two kinds of costs of learning, direct (and measurable) monetary costs for buying information and (unobservable) mental costs of memorizing the information displayed on the screen.

FINAL TEST: Finally, participants took the third *combined knowledge and memory test* in which they had to rank sets of three cities according to the size of their populations. The test is not a pure knowledge test as it includes many smaller cities where a prior pilot has shown that even very knowledgeable subjects may not be able to rank all tuples perfectly without further acquired knowledge from the investment stage. The key idea of the third test is that both, prior knowledge of geography and knowledge acquired during the experiment matter for success. Subjects earned 10 euros if they performed better than the average of participants in the session who

[19]The part of the budget that was not spent, was added to the payoff in the end of the experiment and subjects were aware of this.

[20]This provided them with a task when they finished memorizing or wanted to take a break and induced some opportunity costs of effort.

did the tests in the same order as them.

After the test, participants filled in a questionnaire. In the very end they were informed about how much money they had earned (and how they had performed) in each stage of the experiment.

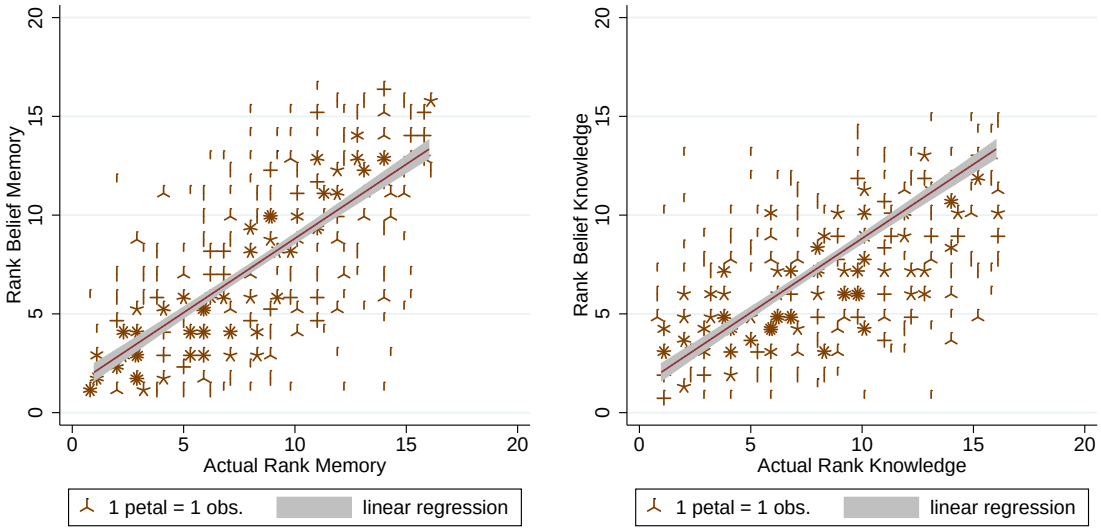
## 5.5 Experimental Results

Our main interest is in the size of the learning investment that participants make to prepare for the final test and how this investment is causally affected by confidence in gains and confidence in levels, i.e. beliefs about learning ability and prior knowledge. The key hypotheses are: (i) confidence in the ability to learn should raise learning efforts irrespective of the prior level of knowledge and (ii) confidence in knowledge should increase the incentives to learn for subjects with low prior knowledge and decrease incentives for subjects with high prior knowledge. We measure confidence as agents' beliefs about their relative rank in the memory and knowledge tests elicited after they have learned the respective feedback scores. We ran 16 experimental sessions in May and June and 8 sessions in October 2015. In total 645 people participated in them.[21] The average total payoff was 11.29 euros (including a 2.50 euros show-up fee), the standard deviation of payoffs was 5.01 euros. Subjects on average earned 1.03 euros in the memory test, 0.89 euros in the knowledge test and 5.00 euros in the final test. Sessions lasted approximately one hour and 10 minutes. 63 percent of participants were female. All participants were university students. The mean semester they were in was 6.5.

[21] Instrumental variable regressions allow us to estimate the causal effect of beliefs on behavior but come along with a substantial loss in statistical power (See, for instance Cameron and Trivedi (2005, section 4.9.3). ) the extent of which is hard to gauge in advance without prior knowledge about the variance in the respective test scores and the outcome variable. For this reason we decided to run additional sessions in October 2015 to collect more observations. We can use 615 observations in our estimates as we have some missing data due to cases in which subjects did not submit their answers.

### 5.5.1 Descriptive Analysis

Figure 5.3: Actual Ranks Versus Rank Beliefs

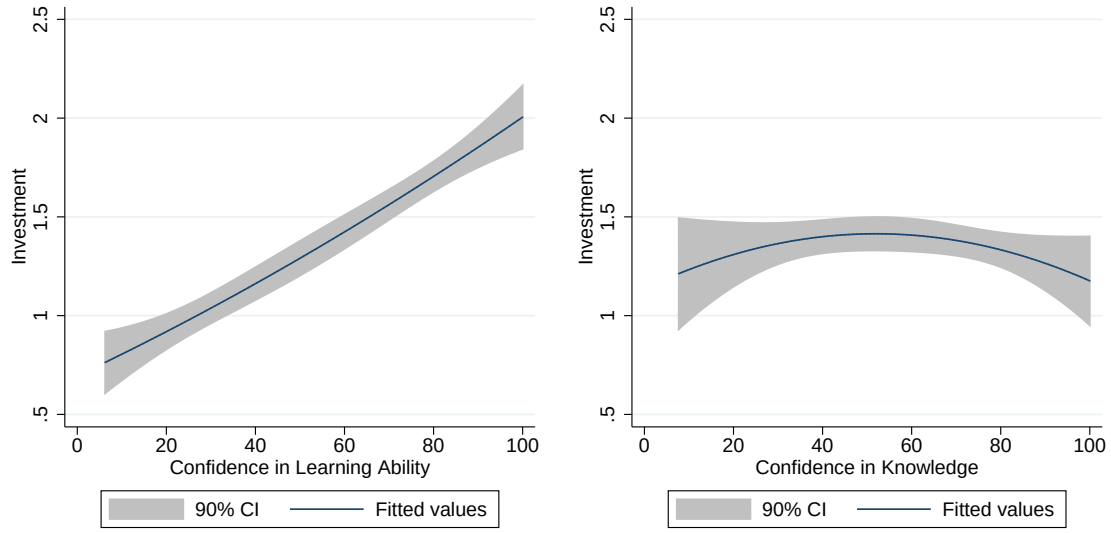



*Note*: This figure shows ordinal ranks versus rank beliefs with respect to the memory and the knowledge test elicited after giving feedback (1 is best).

We begin by descriptively studying the relationship between rank beliefs *elicited after the feedback intervention* and actual ranks as well as the correlation between these beliefs and investment behavior. The sunflower plots in Figure 5.3 show that subjects on average estimate their rank fairly well as most observations are close to the 45 degree line. The correlation of the rank belief in the memory test with the actual rank in the test is 0.75, whereas the correlation of the rank belief in the knowledge test with the actual rank in this test is 0.60. The regression lines in both plots are largely below the 45 degree line indicating that participants on average slightly overestimate their relative performance in both tests (by 0.7 and 1.5 ranks in the memory test and the knowledge test, respectively).[22]

[22]With respect to unincentivized estimates elicited before the feedback intervention, the correlation of beliefs in own performance with one's actual performance (in correct answers) were 0.53 and 0.18 for the memory and the knowledge test, respectively. The correlation between beliefs about one's group's average performance and one's group's actual average performance is 0.10 with respect to the memory test and -0.03 with respect to the knowledge test. Thus, uncertainty was generally high before the intervention, particularly so with respect to others' performance and the prior knowledge dimension. Participants before the intervention were on average slightly underconfident with respect to their own performance (by 0.2 and 1.5 points in the memory and the knowledge test, respectively) and slightly overestimated their group's average performance (by 0.7 and 0.12 points, respectively).

Figure 5.4: Association of Confidence in Learning Ability and in Prior Knowledge with Investment in Learning




*Note*: This figure shows quadratic predictions of learning investment as a function of confidence in learning ability and confidence in prior knowledge.

Figure 5.4 shows quadratic predictions of investment behavior as a function of the respective belief measured in percentile ranks. To facilitate interpretation of coefficients we computed inverted rank beliefs and standardized them to percentile ranks such that the maximum possible level of confidence is 100 and the minimum possible level of confidence is 0.

They thus show the quadratic approximation of the expectation about the level of investment conditional on the two confidence dimensions. As can be seen in the left panel of Figure 5.4, there is a monotonically increasing relationship between confidence in learning ability and monetary investments in learning. The better a person thinks her memory is compared to other people, the larger the amount of costly information she acquires for the study period. The right panel of Figure 5.4 shows that the relationship between the belief in level of prior knowledge and the investment in studying is hump shaped. Investment seems to be the highest if the person thinks that her knowledge is about average.[23]

In the following we will investigate whether these correlations between beliefs and investments are indeed driven by a direct causal effect of beliefs on investments.

[23] A fractional polynomial plot shows nearly exactly the same hump shaped pattern.

In order to do so, we will first check whether our random feedback manipulation affects beliefs as expected. After ensuring that it does, we will use our manipulation to instrument the beliefs in instrumental variable regressions explaining behavior and outcomes. By doing so, we will only use the exogenous component of beliefs, uncorrelated with other unobserved individual traits, to explain behavior.

### 5.5.2 Effect of the Feedback Manipulation on Beliefs

In order to identify the effect of our feedback manipulation on participants' beliefs, we first regress our incentivized measures of confidence in learning ability and confidence in knowledge (i.e. the subjects' beliefs about their respective rank in the considered dimension elicited after the feedback, inverted and standardized to percentile ranks) on the exogenously varied noise terms. We thus estimate the following specification by ordinary least squares, which will also constitute the first stage in our instrumental variable (IV) regressions below:

$$\begin{aligned} Confidence_i = \alpha + \beta NoiseTermMemory_i + \\ \gamma NoiseTermKnowledge_i + \delta Controls_i + \epsilon_i \end{aligned} \tag{5.3}$$

In these, as well as in all of the following regressions, we include the results of the memory and the knowledge test. Additionally, we include dummies for gender, field of study, semester of study, school GPA, income and session as control variables in all regressions.[24] All regressions also include a constant.

The results are reported in Table 1 and show that the respective noise term indeed has a strong effect on the participants' beliefs about their memory and their knowledge. A one unit increase in the noise term in the memory feedback on average causes participants to believe that their memory is 7.6 percentile ranks better whereas a one unit increase in the noise term in the knowledge feedback on average

[24]Tables in appendix 5.7.4 report the regressions without these control variables.

Table 5.1: First Stage Regressions

| | (1) Confidence Memory | (2) Confidence Knowledge |
|---|---|---|
| Noise Term Memory | 7.620*** | -0.142 |
| | (16.30) | (-0.30) |
| Noise Term Knowledge | -0.442 | 5.853*** |
| | (-0.93) | (12.87) |
| Sum Memory Test | 8.564*** | -0.392 |
| | (33.77) | (-1.64) |
| Sum Knowledge Test | -0.453 | 5.886*** |
| | (-1.40) | (17.42) |
| Female | -0.796 | -5.040*** |
| | (-0.55) | (-3.48) |
| $R^2$ | 0.767 | 0.625 |
| Sample Size | 615 | 615 |

*Note*: OLS estimates with robust standard errors; $t$ statistics in parentheses; both regressions contain a constant; additional control variables: dummy variables for gender, field of study (10), semester of study (22), school GPA (25), income (14) and session (24); * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

causes participants to believe their knowledge is 5.9 percentile ranks better. Note that both coefficients have about the same magnitude as the respective coefficients of the true outcomes of the ability tests. Hence, our manipulation worked and the exogenous variation in feedback scores indeed affects beliefs. In the following two subsections, we can now use the manipulation to study the causal effect of confidence in learning ability and prior knowledge on investment behavior and test outcomes.

### 5.5.3 Causal Effect of Beliefs on Learning Investments

By studying whether our treatment affected behavior through affecting beliefs we can address the question of whether the relationships presented in Figure 5.4 indeed reflect causal effects. This will allow us to test the hypotheses stated in section 3. In order to do so, we run an instrumental variable regression of beliefs on investments where the two beliefs are instrumented by the two noise terms. The first stage of the IV regression is given by equation 5.3. As to the second stage, we start by estimating the specification

$$Investment_i = \alpha + \beta ConfidenceMemory_i +$$
$$\gamma ConfidenceKnowledge_i + \delta Controls_i + \epsilon_i \quad (5.4)$$

on the whole sample, including our battery of control variables. Given the hump shaped prediction with respect to the effect of confidence in prior knowledge and the availability of only two instruments, we then split the sample at the median outcome of the knowledge test[25] and estimate effects for the worse half and the better half separately. The results are reported in Table 2.

Table 5.2: Confidence on Investment (IV)

| | (1) Invest. (All) | (2) Invest. (Better) | (3) Invest. (Worse) |
|---|---|---|---|
| Confidence Memory | 0.00792** | 0.00949** | 0.0117*** |
| | (2.42) | (2.00) | (2.67) |
| Confidence Knowledge | -0.00138 | -0.00871* | 0.0147** |
| | (-0.32) | (-1.76) | (2.24) |
| Sum Memory Test | 0.0596* | 0.0212 | 0.0350 |
| | (1.92) | (0.48) | (0.85) |
| Sum Knowledge Test | -0.0159 | -0.0822* | -0.115** |
| | (-0.54) | (-1.94) | (-2.42) |
| Female | 0.00669 | -0.124 | 0.232** |
| | (0.08) | (-1.23) | (2.02) |
| $R^2$ | 0.319 | 0.391 | 0.486 |
| Sample Size | 615 | 353 | 262 |
| F-Test (weak ID), Memory | 136.6 | 56.55 | 52.08 |
| F-Test (weak ID), Knowledge | 83.17 | 54.33 | 26.71 |

*Note*: Two-stage least squares estimates with robust standard errors; $t$ statistics in parentheses; all regressions contain a constant; additional control variables: dummy variables for gender,field of study (10), semester of study (22), school GPA (25), income (14) and session (24); Model 1: whole sample; Model 2: performance at or above the median in knowledge test; Model 3: below median performance in knowledge test; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Column (1) of Table 2 shows that confidence in learning ability significantly increases investment whereas the effect of confidence in levels of prior knowledge is insignificant when looking at the whole sample. Since we expected a positive effect for individuals with low prior knowledge and a negative effect for individuals with high prior knowledge, we split the sample. In columns (2) and (3) we can see that

[25]Median performance was 9 correct sets and we have 119 observation exactly at the median.

both in the better and in the worse half of participants, confidence in learning ability has a positive effect on learning investment. In line with our predictions, we also observe that confidence in levels of knowledge has a negative effect on individuals with above average levels of prior knowledge but a positive effect on individuals with below average levels of prior knowledge. More specifically, for confidence in learning ability we find that an increase of confidence by 10 percentile ranks raises investment in learning by about 9 euro cents for the better half of students and about 12 euro cents for the worse half of students. These effects are significant at the 5 percent and the 1 percent level, respectively. For confidence in knowledge we find that an increase of confidence by 10 percentile ranks lowers investment in learning by about 9 euro cents for students with above average level of prior knowledge but raises investment in learning by about 15 euro cents for students with below average level of prior knowledge. These effects, respectively, are significant at the 10 percent and the 5 percent level. F-tests indicate that our instruments are sufficiently strong.

The experimental results show that beliefs about abilities causally affect how much a person invests in learning. We find that people on average make larger investments in learning the better they believe their learning ability to be. We also find evidence in favor of the hypothesis that increasing the confidence in prior knowledge reduces incentives for individuals whose knowledge is already above average but increases incentives for individuals whose knowledge is below average.

### 5.5.4 Causal Effect of Beliefs on Test Outcomes

We are also interested in whether the behavioral change we brought about by changing confidence beliefs has an effect on students' outcomes in the final test. We begin by estimating how beliefs causally affect the rank one received in the final test. Note that the first stage of the IV regressions is again given by equation 5.3. The second stage is given by:

$$Rank_i = \alpha + \beta ConfidenceMemory_i +$$
$$\gamma ConfidenceKnowledge_i + \delta Controls_i + \epsilon_i \quad (5.5)$$

As can be seen in columns (1) and (2) of Table 3, for the better half of participants in the knowledge test we find no effect of confidence in learning ability[26] but we do find a negative effect of confidence in knowledge again. As confidence in knowledge increases by one percentile rank the outcome in the final test decreases by about 0.3 percentile ranks. For the worse half of participants in the knowledge test we find that as confidence in learning ability increases by one percentile rank the outcome in the final test increases by about 0.3 percentile ranks, while as confidence in prior knowledge increases by one percentile rank the outcome in the final test increases by about 0.5 percentile ranks.

We then use an IV probit estimation method based on Newey (1987) to test whether beliefs also causally affect the probability of passing the test. The first stage is again given by equation 5.3. The second stage is a probit regression of the form

$$Pr(y = 1|x) = G(\alpha + \beta ConfidenceMemory_i +$$
$$\gamma ConfidenceKnowledge_i + \delta Controls + \epsilon_i) \quad (5.6)$$

As can be seen by looking at columns (3) and (4) of Table 3, and analogously to the results in columns (1) and (2), we find that raising the confidence in memory increases the passing probability of people who performed in the worse half in the knowledge test, whereas raising the confidence in prior knowledge decreases the

[26]Hence, for subjects in the better half, the effect of a higher confidence in learning ability on higher learning investments does not translate into better test outcomes. One possible explanation is a physical limitation to the subjects' short term working memory. While more confident subjects were further motivated to acquire knowledge (and thus invested more), they may have been unable to memorize this information appropriately in the given time frame.

passing probability of above average and increases the passing probability of below average performers in the knowledge test. We do not find a significant effect of confidence in memory for individuals who performed in the better half in the knowledge test.

Table 5.3: Confidence on Rank and Probability of Passing Final Test (IV)

| | (1) Rank (Better) | (2) Rank (Worse) | (3) Pr. Pass. (Better) | (4) Pr. Pass. (Worse) |
|---|---|---|---|---|
| Confidence Memory | -0.108 | 0.320** | -0.00106 | 0.0267** |
| | (-0.66) | (2.26) | (-0.12) | (2.13) |
| Confidence Knowledge | -0.297* | 0.549** | -0.0190* | 0.0398** |
| | (-1.68) | (2.36) | (-1.93) | (2.23) |
| Sum Memory Test | 3.478** | 0.722 | 0.111 | -0.0846 |
| | (2.23) | (0.53) | (1.30) | (-0.71) |
| Sum Knowledge Test | 1.071 | -2.137 | 0.0458 | -0.120 |
| | (0.71) | (-1.28) | (0.53) | (-0.92) |
| Female | -10.13*** | 0.496 | -0.502*** | -0.299 |
| | (-3.02) | (0.14) | (-2.64) | (-0.94) |
| $R^2$ | 0.234 | 0.375 | | |
| Sample Size | 353 | 262 | 339 | 235 |
| F-Test (weak ID), M. | 56.55 | 52.08 | | |
| F-Test (weak ID), K. | 54.33 | 26.71 | | |

*Note*: Models 1 and 2: two-stage least squares estimates with robust standard errors; Models 3 and 4: Newey's two-step estimator for binary endogenous variables; $t$ statistics in parentheses; all regressions contain a constant; additional control variables: dummy variables for gender,field of study (10), semester of study (22), school GPA (25), income (14) and session (24); Models 1 and 3: performance at or above the median in knowledge test; Model 2 and 4: below median performance in knowledge test; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

## 5.6 Conclusion

We studied the causal effects of confidence in prior knowledge and in the ability to learn in a lab experiment. Based on a simple formal model, we hypothesized that a higher confidence in one's level of prior knowledge causes students with low levels of knowledge to invest more. This is because it raises the probability that an additional remembered fact is pivotal to passing the test. For students with high levels of prior knowledge we expected the opposite, i.e. that raising their confidence in knowledge would lower their effort to prepare for the test because it subjectively moves them further away from the passing threshold such that learning becomes less relevant

for whether someone passes or fails the test. For the other dimension, confidence in one's learning ability, we expected that raising this dimension of confidence would cause students to invest more effort in learning irrespective of the prior knowledge because the perceived marginal cost of effort decreases.

Our results support these hypotheses. Confidence in learning ability, indeed, raises learning investments irrespective of the prior level of knowledge, whereas confidence in prior knowledge has a negative effect on individuals with above average prior knowledge and a positive effect on individuals with below average prior knowledge on investments. Some of the behavioral effects of our feedback intervention are also reflected by the test outcomes. Raising confidence in learning ability improves the rank and increases the probability of an individual with below average prior knowledge passing the test, whereas we do not find a significant effect for the rank or passing probability of above average individuals. Furthermore, raising confidence in prior knowledge improves the rank and increases the probability that an individual with below average prior knowledge passes the test, whereas it worsens the rank and decreases the passing probability of individuals with above average prior knowledge.

We thus have shown that confidence affects investments in learning in very different ways depending on the specific dimension the belief refers to. People invest more in learning when their confidence in the ability to learn is raised and we find no evidence of a detrimental effect of "too much confidence" in learning ability. Of course, we caution that we studied a lab experiment in a specific content area, and further work has to be done to investigate the validity of the results in other contexts. However, the results already show that generalized statements about the role of confidence can be misleading and confidence should be viewed as a multidimensional

concept.[27]

Insights about the different effects of confidence in learning ability and confidence in prior knowledge have implications not only for the design of interventions aimed at positively affecting academic motivation but also for subjective performance evaluation policies in firms and other organizations. A large literature in psychology and economics has, for instance, stressed that subjective performance evaluations tend to be biased and, in particular, evaluators often tend to be too lenient (see e.g. Murphy and Cleveland 1995; Prendergast 1999). Our results imply that rater leniency (i.e. the tendency to assign too generous performance ratings) can raise motivation when the rater assesses an individual's ability to learn. However, leniency in the rating of a skill level can reduce the motivation as it may signal that one has "already done enough". Hence, while raising confidence in the ability to acquire a certain skill or achieve an outcome can be beneficial, raising confidence in the skill itself or the level of past achievements can be detrimental.

Finally, we note that while we wanted to identify the causal effect of confidence on performance, we did not intend to evaluate the usefulness of confidence *manipulations* in real world settings. The confidence manipulation through noise terms added to test results is designed as a research tool that makes it possible to study causal effects of confidence. It is not meant as an intervention that should be implemented to raise confidence in field settings but we believe that our work can inform the optimal design of interventions that aim at influencing confidence to raise motivation in the field. For instance, our results indicate that interventions that raise the confidence in the ability to learn and grow should be beneficial. Our results are thus well in line with the idea of inducing a "growth mindset", i.e. the belief that intelligence is malleable rather than fixed, which has been shown to raise educational

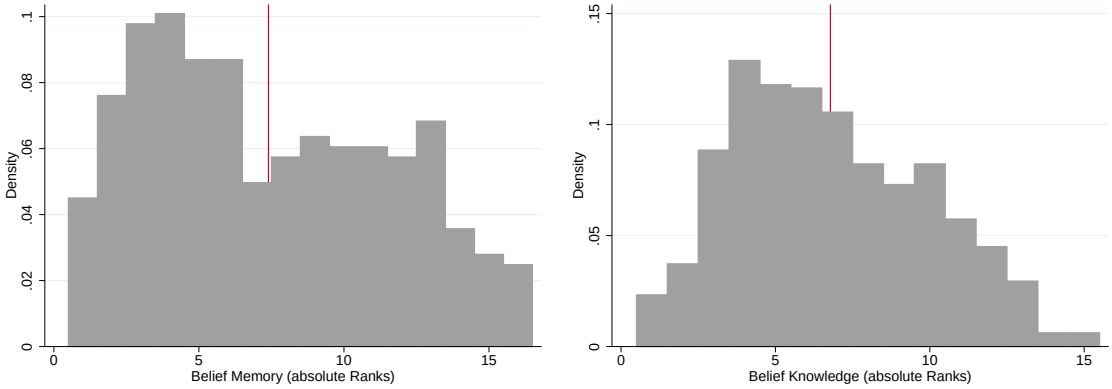
[27]Interestingly, we find that women are significantly less confident than men with respect to their prior knowledge (skill level) but not so with respect to their memory (ability to acquire new skills). This further hints towards the importance of a multidimensional understanding of confidence for explaining gender effects in competitive settings. In settings where skill level is important, women are observed to shy away from competition, partly due to lower confidence (Niederle and Vesterlund, 2007). It should be further explored what happens in settings where beliefs about the ability to learn play a role.

outcomes (Yeager et al., 2014; Paunesku et al., 2015; Alan et al., 2016). However, our results also show that interventions that raise confidence in traits that directly contribute to outcomes (such a prior knowledge) may be detrimental.

## 5.7 Appendix to Chapter 5

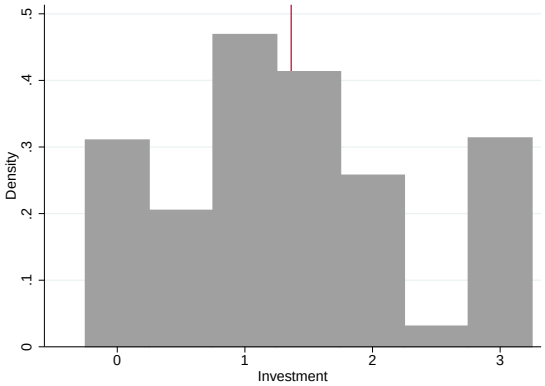
### 5.7.1 Descriptive Statistics and Figures

Figure 5.5: Rank Beliefs




*Note*: Distributions and means of rank beliefs elicited after giving feedback. (1 is best)

Figure 5.6: Investment (in Euros)




*Note*: Distribution of learning investments in euros.

Table 5.4: Summary Statistics

| Variable | Mean | Std. Dev. | Min. | Max. | N |
|---|---|---|---|---|---|
| Noise Term Memory | -0.03 | 1.42 | -2 | 2 | 644 |
| Noise Term Knowledge | 0.01 | 1.4 | -2 | 2 | 644 |
| Belief Memory | 54 | 28.54 | 6.25 | 100 | 644 |
| Belief Knowledge | 58.32 | 21.51 | 7.69 | 100 | 644 |
| Sum Memory Test | 5.15 | 2.55 | 0 | 11 | 644 |
| Sum Knowledge Test | 8.87 | 2.18 | 0 | 16 | 644 |
| Sum Test 3 | 10.72 | 2.52 | 1 | 20 | 644 |
| Investment | 1.36 | 0.95 | 0 | 3 | 644 |
| Prob. of Passing Test 3 | 0.5 | 0.5 | 0 | 1 | 644 |
| Profit | 11.29 | 5.02 | 3.2 | 19.4 | 644 |
| Female | 0.63 | 0.48 | 0 | 1 | 644 |
| School GPA | 2.05 | 0.6 | 1 | 3.5 | 623 |
| Humanities | 0.16 | 0.37 | 0 | 1 | 644 |
| Social Sciences | 0.09 | 0.29 | 0 | 1 | 644 |
| Law | 0.05 | 0.22 | 0 | 1 | 644 |
| Business | 0.26 | 0.44 | 0 | 1 | 644 |
| Economics | 0.13 | 0.34 | 0 | 1 | 644 |
| Medicine | 0.05 | 0.21 | 0 | 1 | 644 |
| Natural Sciences | 0.08 | 0.27 | 0 | 1 | 644 |
| Psychology | 0.01 | 0.12 | 0 | 1 | 644 |
| Other Subjects | 0.14 | 0.35 | 0 | 1 | 644 |
| Non-Student | 0.02 | 0.13 | 0 | 1 | 644 |
| Semester 1 | 0.06 | 0.24 | 0 | 1 | 635 |
| Semester 2 | 0.11 | 0.31 | 0 | 1 | 635 |
| Semester 3 | 0.06 | 0.23 | 0 | 1 | 635 |
| Semester 4 | 0.12 | 0.33 | 0 | 1 | 635 |
| Semester 5 | 0.08 | 0.27 | 0 | 1 | 635 |
| Semester 6 | 0.13 | 0.33 | 0 | 1 | 635 |
| Semester 7 | 0.08 | 0.27 | 0 | 1 | 635 |
| Semester 8 | 0.1 | 0.31 | 0 | 1 | 635 |
| Semester 9 | 0.06 | 0.24 | 0 | 1 | 635 |
| Semester 10 | 0.05 | 0.22 | 0 | 1 | 635 |
| Semester 11 | 0.04 | 0.2 | 0 | 1 | 635 |
| Semester 12 | 0.03 | 0.18 | 0 | 1 | 635 |
| Semester 13 | 0.02 | 0.14 | 0 | 1 | 635 |
| Semester 14 | 0.01 | 0.11 | 0 | 1 | 635 |
| Semester 15 | 0.01 | 0.1 | 0 | 1 | 635 |
| Semester 16 | 0.01 | 0.1 | 0 | 1 | 635 |
| Semester 17 | 0 | 0.06 | 0 | 1 | 635 |
| Semester 18 | 0.01 | 0.08 | 0 | 1 | 635 |
| Semester 19 | 0 | 0.04 | 0 | 1 | 635 |
| Semester 20 | 0 | 0.04 | 0 | 1 | 635 |
| Semester 21 | 0 | 0.06 | 0 | 1 | 635 |
| Semester 23 | 0 | 0.04 | 0 | 1 | 635 |
| Session 1 | 0.04 | 0.19 | 0 | 1 | 644 |
| Session 2 | 0.04 | 0.19 | 0 | 1 | 644 |
| Session 3 | 0.04 | 0.2 | 0 | 1 | 644 |
| Session 4 | 0.03 | 0.17 | 0 | 1 | 644 |
| Session 5 | 0.04 | 0.19 | 0 | 1 | 644 |
| Session 6 | 0.05 | 0.21 | 0 | 1 | 644 |
| Session 7 | 0.03 | 0.18 | 0 | 1 | 644 |
| Session 8 | 0.04 | 0.19 | 0 | 1 | 644 |
| Session 9 | 0.05 | 0.21 | 0 | 1 | 644 |
| Session 10 | 0.05 | 0.21 | 0 | 1 | 644 |
| Session 11 | 0.05 | 0.22 | 0 | 1 | 644 |
| Session 12 | 0.03 | 0.17 | 0 | 1 | 644 |
| Session 13 | 0.05 | 0.22 | 0 | 1 | 644 |
| Session 14 | 0.04 | 0.2 | 0 | 1 | 644 |
| Session 15 | 0.03 | 0.16 | 0 | 1 | 644 |
| Session 16 | 0.03 | 0.18 | 0 | 1 | 644 |
| Session 17 | 0.05 | 0.21 | 0 | 1 | 644 |
| Session 18 | 0.05 | 0.22 | 0 | 1 | 644 |
| Session 19 | 0.05 | 0.22 | 0 | 1 | 644 |
| Session 20 | 0.05 | 0.21 | 0 | 1 | 644 |
| Session 21 | 0.04 | 0.2 | 0 | 1 | 644 |
| Session 22 | 0.05 | 0.22 | 0 | 1 | 644 |
| Session 23 | 0.04 | 0.2 | 0 | 1 | 644 |
| Session 24 | 0.05 | 0.21 | 0 | 1 | 644 |

### 5.7.2 OLS Regressions of Beliefs on Behavior and Outcomes

Table 5.5: Confidence on Investment (OLS)

| | (1) Invest. (All) | (2) Invest. (Better) | (3) Invest. (Worse) |
|---|---|---|---|
| Belief Memory | 0.00871*** | 0.00637** | 0.0120*** |
| | (3.97) | (2.03) | (3.38) |
| Belief Knowledge | 0.00239 | 0.00234 | 0.000527 |
| | (1.11) | (0.78) | (0.15) |
| Sum Memory Test | 0.0552** | 0.0550 | 0.0189 |
| | (2.15) | (1.53) | (0.47) |
| Sum Knowledge Test | -0.0380* | -0.146*** | -0.0379 |
| | (-1.84) | (-3.47) | (-0.89) |
| Female | 0.0264 | -0.0559 | 0.205 |
| | (0.32) | (-0.50) | (1.51) |
| $R^2$ | 0.323 | 0.421 | 0.528 |
| Sample Size | 615 | 353 | 262 |

*Note*: OLS estimates with robust standard errors; t statistics in parentheses; all regressions contain a constant; additional control variables: dummy variables for gender, field of study (10), semester of study (22), school GPA (25), income (14) and session (24); * p<0.10, ** p<0.05, *** p<0.01.

Table 5.6: Confidence on Outcomes (OLS)

| | (1) Rank (Better) | (2) Rank (Worse) | (3) Prob. Pass. (Better) | (4) Prob. Pass. (Worse) |
|---|---|---|---|---|
| Belief Memory | -0.0501 | 0.103 | -0.0000928 | 0.00961 |
| | (-0.50) | (1.00) | (-0.02) | (1.51) |
| Belief Knowledge | -0.0158 | 0.283** | -0.00992* | 0.0210*** |
| | (-0.14) | (2.15) | (-1.91) | (2.79) |
| Sum Memory Test | 3.149*** | 2.368** | 0.107** | 0.0472 |
| | (2.83) | (2.09) | (2.03) | (0.70) |
| Sum Knowledge Test | -0.481 | -1.090 | -0.00528 | -0.0569 |
| | (-0.32) | (-0.65) | (-0.08) | (-0.66) |
| Female | -8.536** | 0.786 | -0.446** | -0.222 |
| | (-2.23) | (0.19) | (-2.45) | (-0.87) |
| $R^2$ | 0.254 | 0.405 | | |
| $Chi^2$ | | | 84.89 | 137.9 |
| Sample Size | 353 | 262 | 339 | 235 |

*Note*: OLS estimates with robust standard errors; t statistics in parentheses; all regressions contain a constant; additional control variables: dummy variables for gender, field of study (10), semester of study (22), school GPA (25), income (14) and session (24); * p<0.10, ** p<0.05, *** p<0.01.

### 5.7.3 Reduced Form Estimates

Table 5.7: Noise Terms on Investment (OLS)

| | (1) Invest. (All) (OLS) | (2) Invest. (Better) (OLS) | (3) Invest. (Worse)(OLS) |
|---|---|---|---|
| Noise Term Memory | 0.0606** | 0.0660* | 0.0917** |
| | (2.21) | (1.74) | (2.01) |
| Noise Term Knowledge | -0.0116 | -0.0534 | 0.0797* |
| | (-0.42) | (-1.53) | (1.84) |
| Sum Memory Test | 0.128*** | 0.103*** | 0.125*** |
| | (8.35) | (4.90) | (4.96) |
| Sum Knowledge Test | -0.0276* | -0.136*** | -0.0572 |
| | (-1.65) | (-3.65) | (-1.34) |
| Female | 0.00732 | -0.0847 | 0.231* |
| | (0.09) | (-0.77) | (1.71) |
| $R^2$ | 0.305 | 0.423 | 0.506 |
| Sample Size | 615 | 353 | 262 |

*Note*: OLS estimates with robust standard errors; t statistics in parentheses; all regressions contain a constant; additional control variables: dummy variables for gender, field of study (10), semester of study (22), school GPA (25), income (14) and session (24); * p<0.10, ** p<0.05, *** p<0.01.

Table 5.8: Noise Terms on Outcomes (OLS)

| | (1) Rank (Better) (OLS) | (2) Rank (Worse)(OLS) | (3) Prob. Win. (Better) (Probit) | (4) Prob. Win. (Worse)(Probit) |
|---|---|---|---|---|
| main | | | | |
| Noise Term Memory | -0.703 | 2.408* | -0.00914 | 0.182** |
| | (-0.55) | (1.66) | (-0.15) | (2.17) |
| Noise Term Knowledge | -1.861 | 2.979** | -0.118* | 0.219*** |
| | (-1.46) | (1.98) | (-1.96) | (2.77) |
| Sum Memory Test | 2.679*** | 3.065*** | 0.109*** | 0.118*** |
| | (3.74) | (4.08) | (3.16) | (2.92) |
| Sum Knowledge Test | -0.587 | 0.168 | -0.0589 | 0.0237 |
| | (-0.44) | (0.10) | (-0.94) | (0.29) |
| Female | -8.407** | 0.348 | -0.401** | -0.302 |
| | (-2.19) | (0.08) | (-2.19) | (-1.21) |
| $R^2$ | 0.261 | 0.404 | | |
| $Chi^2$ | | | 83.82 | 133.6 |
| Sample Size | 353 | 262 | 339 | 235 |

*Note*: OLS estimates with robust standard errors; t statistics in parentheses; all regressions contain a constant; additional control variables: dummy variables for gender, field of study (10), semester of study (22), school GPA (25), income (14) and session (24); * p<0.10, ** p<0.05, *** p<0.01.

### 5.7.4 Results Without Session Dummies and Demographic Control Variables

Table 5.9: First Stage Regressions Without Additional Control Variables

| | (1) Confidence Memory | (2) Confidence Knowledge |
|---|---|---|
| Noise Term Memory | 7.468*** | -0.582 |
| | (18.04) | (-1.39) |
| Noise Term Knowledge | -0.684* | 6.239*** |
| | (-1.70) | (14.96) |
| Sum Memory Test | 8.551*** | -0.204 |
| | (34.89) | (-0.87) |
| Sum Knowledge Test | -0.412 | 5.909*** |
| | (-1.40) | (19.71) |
| Constant | 13.88*** | 6.864** |
| | (4.50) | (2.36) |
| $R^2$ | 0.727 | 0.530 |
| Sample Size | 644 | 644 |

*Note*: OLS estimates with robust standard errors; $t$ statistics in parentheses;* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5.10: Confidence on Investment (IV) Without Additional Control Variables

| | (1) Invest. (All) | (2) Invest. (Better) | (3) Invest. (Worse) |
|---|---|---|---|
| Confidence Memory | 0.00906*** | 0.00646 | 0.0118** |
| | (2.67) | (1.34) | (2.42) |
| Confidence Knowledge | -0.000832 | -0.00581 | 0.00627 |
| | (-0.21) | (-1.14) | (0.91) |
| Sum Memory Test | 0.0531* | 0.0718 | 0.0373 |
| | (1.68) | (1.61) | (0.83) |
| Sum Knowledge Test | -0.00194 | -0.0467 | -0.0662 |
| | (-0.07) | (-1.07) | (-1.37) |
| Constant | 0.666*** | 1.557*** | 0.654** |
| | (4.05) | (3.98) | (2.16) |
| $R^2$ | 0.164 | 0.137 | 0.175 |
| Sample Size | 644 | 374 | 270 |
| F-Test (weak ID), Memory | 164.4 | 87.01 | 80.77 |
| F-Test (weak ID), Knowledge | 112.4 | 78.50 | 36.53 |

*Note*: Two-stage least squares estimates with robust standard errors; $t$ statistics in parentheses; Model 1: whole sample; Model 2: performance at or above the median in knowledge test; Model 3: below median performance in knowledge test; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5.11: Confidence on Rank and Probability of Passing Final Test (IV) Without Additional Control Variables

| | (1) Rank (Better) | (2) Rank (Worse) | (3) Pr. Pass. (Better) | (4) Pr. Pass. (Worse) |
|---|---|---|---|---|
| Confidence Memory | -0.114 | 0.258* | -0.00288 | 0.00704 |
| | (-0.76) | (1.74) | (-0.42) | (0.99) |
| Confidence Knowledge | -0.199 | 0.370* | -0.0128* | 0.0151 |
| | (-1.25) | (1.69) | (-1.71) | (1.49) |
| Sum Memory Test | 3.750*** | 1.024 | 0.126* | 0.0297 |
| | (2.70) | (0.72) | (1.93) | (0.45) |
| Sum Knowledge Test | 0.973 | -0.794 | 0.0215 | -0.0573 |
| | (0.68) | (-0.50) | (0.34) | (-0.77) |
| Constant | 51.81*** | 24.15** | 0.261 | -0.989** |
| | (3.85) | (2.24) | (0.45) | (-2.02) |
| $R^2$ | 0.0334 | 0.0756 | | |
| Sample Size | 374 | 270 | 374 | 270 |
| F-Test (weak ID), M. | 87.01 | 80.77 | | |
| F-Test (weak ID), K. | 78.50 | 36.53 | | |

*Note*: Two-stage least squares estimates with robust standard errors; $t$ statistics in parentheses; Model 1: whole sample; Model 2: performance at or above the median in knowledge test; Model 3: below median performance in knowledge test; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

### 5.7.5 Timeline of the Experiment

1. Measurement of Prior Knowledge and Learning Ability: Subjects take two tests (incentivized with piece rate, the order is randomized to control for ordering effects):

    - „knowledge test“: participants have to solve 20 sets of three German cities each by indicating which is the largest, which is the second largest and which is the third largest in terms of population within each triple
    - „memory test“: participants for 15 minutes see a screen with a list of 36 German cities with (arbitrary) four digit „cities codes“ which they can memorize, then they have to solve 12 sets of three cities each by indicating which one has the largest, which one has the second largest, and which one has the third largest city code
    - Immediately after each test participants estimate their number of correct sets and other‘s average number of correct sets in each test (belief elicitation, unincentivized)

2. Information on further course (introduction of combined test): Subjects are informed that there will be a third test and that they earn a prize if their outcome is above average. They are explained how they can prepare for it. Furthermore, they are told that they will receive feedback and given an explanation of how the feedback is computed.

3. Feedback stage: Subjects receive noisy feedback about their performance in both tests (treatment variation)

4. Measurement of confidence (belief elicitation, both tests, incentivized): Subjects estimate their rank in both tests

5. Investment stage (information acquisition): Subjects receive a budget of 3 euros from which they can buy information on cities in increments of 0.5 euros or 10 cities (behavioral outcome variable)

6. Measurement of outcomes (combined knowledge and memory test): Subjects take the third test (economic outcome variables)

### 5.7.6 Details on the Tests, Feedback, Elicitation of Beliefs, and Investment Stage

The experiment was conducted in German, so in the following we give the English translation of the texts. All the cities used in the experiment come from the set of the 200 largest cities in Germany. We pretested all instructions and tests to ensure that they are understandable and produced a sufficient variance of results so that relative performance/ability could be measured precisely. Before the tests started, an introductory screen described the test and how money could be earned. We also made sure that subjects understood the rules of the tests by including a sample exercise before each test and subjects could only start the test after answering it according to the rules.

#### Description of Knowledge Test

The instruction on the introductory screen to the knowledge test said:

> "In the following you can earn money by ordering three cities, respectively, according to their number of inhabitants. In total there are 20 sets of 3 cities each. For each completely correct set you will receive 0.10 euros. If the set was not answered completely correctly you will not receive any money for it. You have 6 minutes to work on the test. Write a 1 in the field next to the city you belief is the largest of the three, write a 2 in the field of the intermediate city and write a 3 in the field next to the smallest city."

On the test screen itself a summary of the instructions and the payment scheme was given. A countdown clock was shown. For example, a set of three cities looked like this:

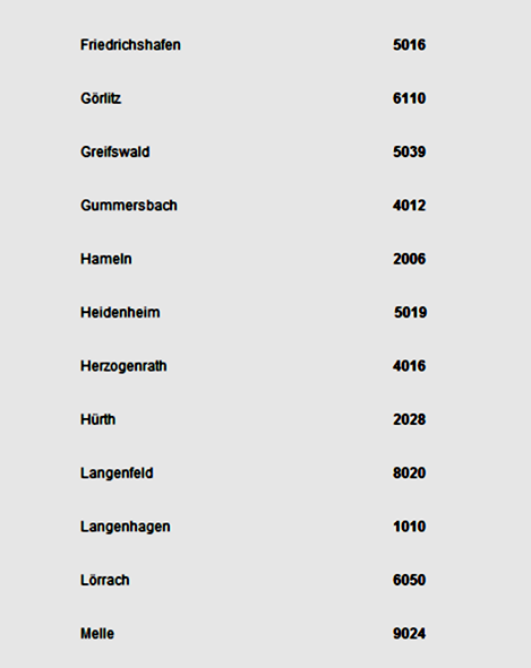

### Description of Memory Test

The instruction on the introductory screen to the memory test said:

> "In the following you can earn money by ordering three cities, respectively, according to their city codes. In total there are 12 sets of 3 cities each. For each completely correct set you will receive 0.20 euros. If the set was not answered completely correctly you will not receive any money for it. You have 6 minutes to work on the test.
>
> Since the city codes are generally not known, you will receive an alphabetically ordered list with all 36 cities and their respective city codes. This list will be displayed to you in a learning phase of 15 minutes. You have the opportunity to memorize the ranking (relative size) of these city codes, in order to later order three cities each according to this number. During the test this list will not be displayed anymore, so that only your memory will help you to do the ordering. Note-taking is not allowed.Violation of this rule will lead to the exclusion from this and future experiments.
>
> Write a 1 in the field next to the city which according to your memory has the largest city code, write a 2 in the field of the city with the second largest city code and write a 3 in the field next to the city with the smallest city code."

On the learning and test screens a summary of the instructions and the payment scheme was given. A countdown clock was shown. The sets of three cities in the memory test looked the same as in the knowledge test but none of the city names

were used twice. Information displayed in the learning phase looked like this:



| | |
|---|---|
| Friedrichshafen | 5016 |
| Görlitz | 6110 |
| Greifswald | 5039 |
| Gummersbach | 4012 |
| Hameln | 2006 |
| Heidenheim | 5019 |
| Herzogenrath | 4016 |
| Hürth | 2028 |
| Langenfeld | 8020 |
| Langenhagen | 1010 |
| Lörrach | 6050 |
| Melle | 9024 |

### Description of Feedback

After subjects have been told that there will be a third "main test" and that they can prepare for it, they are informed that they are about to receive feedback. Next, they are shown a screen where the computation of the "feedback scores" is explained:

> "The experimental software will now generate a knowledge score and a memory score for each participant. The knowledge score is being computed based on a participant's number of correct answers in the knowledge test whereas the memory score is computed based on a participant's number of correct answers in the in the memory test. In expectation, each score is equal to the participant's actual number of correct answers. The experimental software will soon let you know your score.
>
> **Computation of the feedback scores**:
>
> Your scores are composed of the following:
>
> Knowledge score = number of your correct sets in the knowledge test

+ random variable X

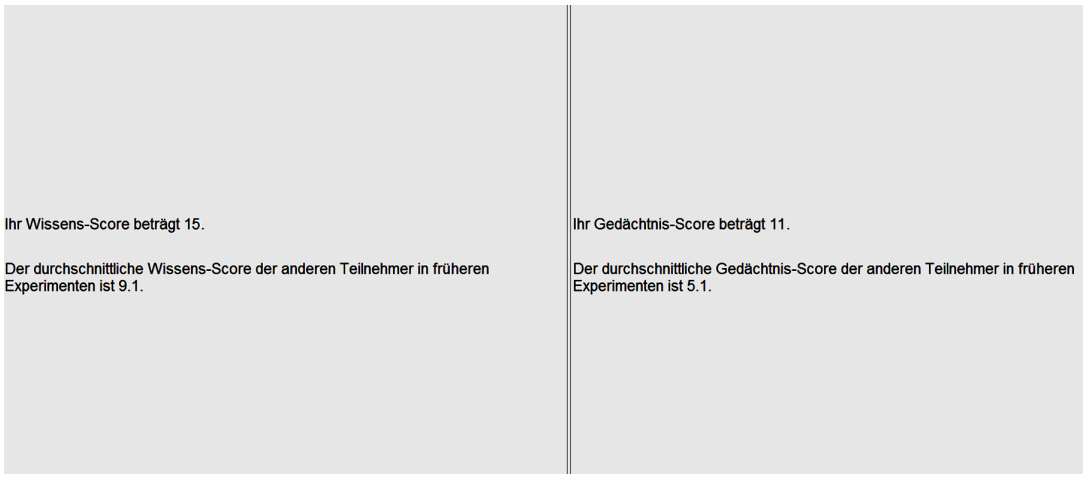
Memory Score = number of your correct sets in the memory test + random variable Y

The random variables X and Y can each assume values between -2 and +2, that means each of the values (-2, -1, 0, +1, +2) is equally likely (i.e. occurs with a probability of 20%). Furthermore, the random variables X and Y are independent of each other, that means also all combinations of values of the random variables X and Y are equally likely."

On the Next screen, subjects receive the following information:

"The knowledge score can help you to assess your knowledge of cities relative to other participants whereas the memory score can help you to assess your memory capacity relative to other participants. The two scores give your number of correct sets in each test with a certain imprecision but in expectation equal the actual number of your correct answers."

The feedback screen displayed both a participant's two scores and the respective average score of participants in earlier experimental sessions: "Your [knowledge/memory] score is [x]. The average [knowledge/memory] score of the other participants in earlier experiments is [9.1/5.1]" It looked like this:



**Elicitation of Beliefs**

The elicitation screen contained the following text:

> "Half of participants in this room worked on the two tests in the same order as you. How do you assess your own results in both tests relative to theses participants? Please estimate your rank below. For each estimate you will earn one euro if you guess the rank exactly right. There are [x] participants in your group.
>
> The participant with the highest number of points occupies rank 1, the participant with the lowest number of points occupies rank [x]."

Then participants could indicate their rank beliefs in the knowledge and the memory test by selecting a number on two lines of radio buttons. The number of radio buttons was automatically adjusted to the number of people in each of the two groups per session.

**Investment Stage**

The decision screen contained the following information:

> "**Description of test 3: combined knowledge and memory test**
>
> In the following you can earn money by ordering three cities, respectively, according to their numbers of inhabitants. In total there are 20 sets of 3 cities each. You have 6 minutes to work on the test.
>
> The cities are German cities of comparable size and prominence as the cities in the knowledge test about the numbers of inhabitants. However, no of these cities will be in the test again.
>
> If your result is above average, that is if you get more correct answers than the average of the participants in the room who worked on the first two tests in the same order as you, you will receive 10 euros, if not you will receive zero euros.

You have the possibility to improve you knowledge of the cities in a learning phase.

**Description of preparation for test 3**

In order to prepare for test 3, you may buy information about cities' numbers of inhabitants. In order to do so you receive, independently of your performance until now, a budget of 3.00 euros. The part of the budget that you do not spend, will be added to your payoff in the end of the experiment. All cities you can buy are part of the test. You can buy packages of 10 cities each. Each package allows you to completely answer at least 3 assignments (sets).

Example for information you can buy:

Innsbruck 121,329

Following your selection, for 15 minutes the program will show in alphabetical order your acquired packages of cities with their respective numbers of inhabitants. This information you may memorize so that you can better order cities according to their size in the main test. Note-taking is not allowed. Violation of this rule will lead to the exclusion from this and future experiments."

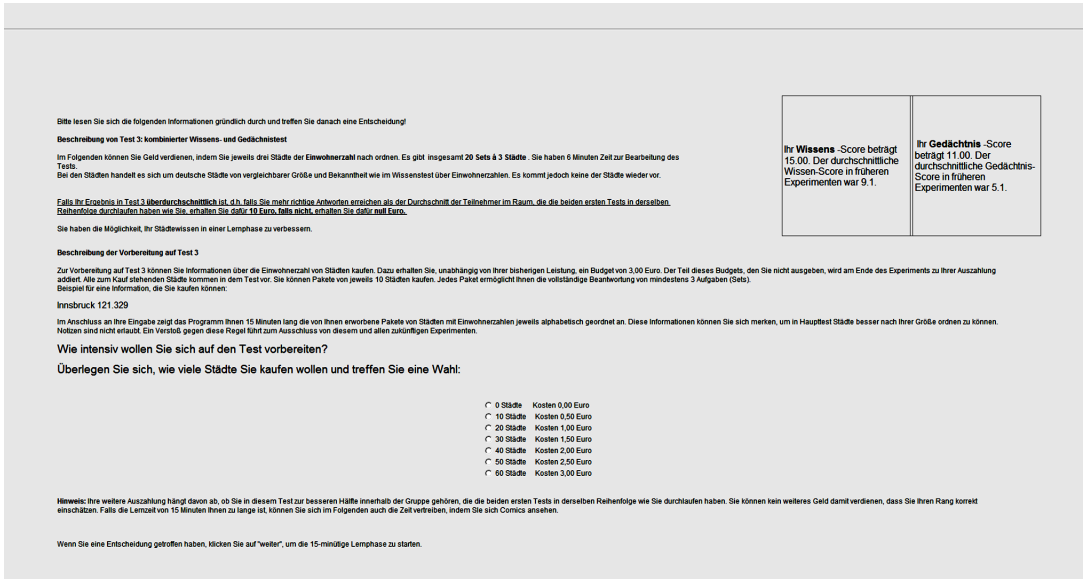
Below this text, subjects were asked to decide how many cities they want to buy and indicate their choice with the respective radio button. They have to make a choice between buying 0, 10, 20, 30, 40, 50, or 60 cities. Each ten cities cost 0.5 euros.

Below the radio buttons it said:

"Please note: Your further payoff depends on whether you belong to the better half of the group who worked on the first two tests in the same order as you. You cannot earn additional money by estimating your rank correctly. In case you find the study time of 15 minutes too long, you can also spend time looking at comics."

A reminder of their knowledge and memory score is displayed in the upper right corner of the screen.

This is how the screen looked like:



Bitte lesen Sie sich die folgenden Informationen gründlich durch und treffen Sie danach eine Entscheidung!

**Beschreibung von Test 3: kombinierter Wissens- und Gedächnistest**

Im Folgenden können Sie Geld verdienen, indem Sie jeweils drei Städte der **Einwohnerzahl** nach ordnen. Es gibt insgesamt **20 Sets à 3 Städte** . Sie haben 6 Minuten Zeit zur Bearbeitung des Tests.
Bei den Städten handelt es sich um deutsche Städte von vergleichbarer Größe und Bekanntheit wie im Wissenstest über Einwohnerzahlen. Es kommt jedoch keine der Städte wieder vor.

Falls Ihr Ergebnis in Test 3 **überdurchschnittlich** ist, d.h. falls Sie mehr richtige Antworten erreichen als der Durchschnitt der Teilnehmer im Raum, die die beiden ersten Tests in derselben Reihenfolge durchlaufen haben wie Sie, erhalten Sie dafür **10 Euro, falls nicht,** erhalten Sie dafür **null Euro.**

Sie haben die Möglichkeit, Ihr Städtewissen in einer Lernphase zu verbessern.

**Beschreibung der Vorbereitung auf Test 3**

Zur Vorbereitung auf Test 3 können Sie Informationen über die Einwohnerzahl von Städten kaufen. Dazu erhalten Sie, unabhängig von Ihrer bisherigen Leistung, ein Budget von 3,00 Euro. Der Teil dieses Budgets, den Sie nicht ausgeben, wird am Ende des Experiments zu Ihrer Auszahlung addiert. Alle zum Kauf stehenden Städte kommen in dem Test vor. Sie können Pakete von jeweils 10 Städten kaufen. Jedes Paket ermöglicht Ihnen die vollständige Beantwortung von mindestens 3 Aufgaben (Sets).
Beispiel für eine Information, die Sie kaufen können:

Innsbruck 121.329

Im Anschluss an Ihre Eingabe zeigt das Programm Ihnen 15 Minuten lang die von Ihnen erworbene Pakete von Städten mit Einwohnerzahlen jeweils alphabetisch geordnet an. Diese Informationen können Sie sich merken, um in Haupttest Städte besser nach Ihrer Größe ordnen zu können. Notizen sind nicht erlaubt. Ein Verstoß gegen diese Regel führt zum Ausschluss von diesem und allen zukünftigen Experimenten.

**Wie intensiv wollen Sie sich auf den Test vorbereiten?**

**Überlegen Sie sich, wie viele Städte Sie kaufen wollen und treffen Sie eine Wahl:**

- 0 Städte Kosten 0,00 Euro
- 10 Städte Kosten 0,50 Euro
- 20 Städte Kosten 1,00 Euro
- 30 Städte Kosten 1,50 Euro
- 40 Städte Kosten 2,00 Euro
- 50 Städte Kosten 2,50 Euro
- 60 Städte Kosten 3,00 Euro

**Hinweis:** Ihre weitere Auszahlung hängt davon ab, ob Sie in diesem Test zur besseren Hälfte innerhalb der Gruppe gehören, die die beiden ersten Tests in derselben Reihenfolge wie Sie durchlaufen haben. Sie können kein weiteres Geld damit verdienen, dass Sie Ihren Rang korrekt einschätzen. Falls die Lernzeit von 15 Minuten Ihnen zu lange ist, können Sie sich im Folgenden auch die Zeit vertreiben, indem Sie sich Comics ansehen.

Wenn Sie eine Entscheidung getroffen haben, klicken Sie auf "weiter", um die 15-minütige Lernphase zu starten.

Ihr **Wissens** -Score beträgt 15.00. Der durchschnittliche Wissen-Score in früheren Experimenten war 9.1.

Ihr **Gedächtnis** -Score beträgt 11.00. Der durchschnittliche Gedächtnis-Score in früheren Experimenten war 5.1.

## Description of Test 3

Test 3 looked the same as the first two tests and contained 20 sets of three cities each. Within each set participants had to order cities according to their numbers of inhabitants. A summary of the instructions and the payment scheme was given.

# Bibliography

Abbott, A. and Leslie, D. (2004). Recent trends in higher education applications and acceptances. *Education Economics*, 12(1):67–86.

Ackerman, P. L., Beier, M. E., and Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs?. *Psychological Bulletin*, 131(1):30–60.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Akerlof, G. A. and Kranton, R. E. (2002). Identity and schooling: Some lessons for the economics of education. *Journal of Economic Literature*, 40(4):1167–1201.

Alan, S., Boneva, T., and Ertac, S. (2016). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *HCEO Working Paper*.

Albrecht, K., Parys, J., Szech, N., and von Essen, E. (2013). Updating, self-confidence and discrimination. *European Economic Review*, pages 144–169.

Alloway, T. P. and Alloway, R. G. (2010). Investigating the predictive roles of working memory and iq in academic attainment. *Journal of Experimental Child Psychology*, 106(1):20–29.

Alos-Ferrer, C. and Strack, F. (2014). From dual processes to multiple selves: Implications for economic behavior. *Journal of Economic Psychology*, 41:1 – 11. From Dual Processes to Multiple Selves: Implications for Economic Behavior.

Angrist, J., Bettinger, E., and Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in colombia. *American Economic Review*, 96(3):847–862.

Angrist, J. and Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, 99(4):1384–1414.

Angrist, J. D. and Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton NJ.

Aoyagi, M. (2010). Information feedback in a dynamic tournament. *Games and Economic Behavior*, 70(2):242–260.

Arrow, K. J. (1973). The theory of discrimination. In Ashenfelter, O. and Rees, A., editors, *Discrimination in Labor Markets*. Princeton University Press, Princeton, N.J.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44 – 63.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2016). What you don't know... Can't hurt you? A field experiment on relative performance feedback in higher education. Discussion Paper DP11201, Centre for Economic Policy Research.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452.

Azmat, G. and Iriberri, N. (2016). The Provision of Relative Performance Feedback: An Analysis of Performance and Satisfaction. *Journal of Economics & Management Strategy*, 25(1):77–110.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13–25.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology*, 4(3):359–373.

Barankay, I. (2012). Rank incentives - evidence from a randomized workplace experiment. *unpublished working paper*.

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292.

Bell, R. and McCaffrey, D. (2002). Bias Reduction in Standard Errors for Linear and Generalized Linear Models with Multi-stage Samples. *Survey Methodology*, 28:169–179.

Benabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Benabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, 70(3):489–520.

Benabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.

Benartzi, S. and Thaler, R. (2007). Heuristics and biases in retirement savings behavior. *Journal of Economic Perspectives*, 21(3):81–104.

Bergh, A. and Fink, G. (2009). Higher education, elite institutions and inequality. *European Economic Review*, 53(3):376 – 384.

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

Black, D., Smith, J., and Daniel, K. (2005). College quality and wages in the united states. *German Economic Review*, 6(3):415–443.

Blanes i Vidal, J. and Nossol, M. (2011). Tournaments without prizes: Evidence from personnel records. *Management Science*, 57(10):1721–1736.

Booij, A. S., Leuven, E., and Oosterbeek, H. (2017). Ability peer effects in university: Evidence from a randomized experiment. *The Review of Economic Studies*, 84(2):547–578.

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2016a). Employee recognition and performance: A field experiment. *Management Science*, 62(11):3085–3099.

Bradler, C., Neckermann, S., and Warnke, A. J. (2016b). Incentivizing creativity: A large-scale experiment with tournaments and gifts. *ZEW Discussion Papers*, 16(040).

Braun, S., Dwenger, N., and Kübler, D. (2010). Telling the truth may not pay off: An empirical study of centralised university admissions in germany. *The B.E. Journal of Economic Analysis and Policy*, 10(1):Article 22.

Bridgeland, J. M., DiIulio, J. J., and Burke Morison, K. (2006). The silent epidemic: Perspectives of high school dropouts. Technical report, Bill & Melinda Gates Foundation.

Broecke, S. (2015). University rankings: do they matter in the uk? *Education Economics*, 23(2):137–161.

Bruckmeier, K., Fischer, G.-B., and Wigger, B. U. (2013). Does Distance Matter? Tuition Fees and Enrollment of First-Year Students at German Public Universities. CESifo Working Paper Series 4258, CESifo Group Munich.

Bruckmeier, K., Fischer, G.-B., and Wigger, B. U. (2014). The Downside Risk of Elevation. CESifo Working Paper Series 4950, CESifo Group Munich.

Bruckmeier, K. and Wigger, B. U. (2014). The effects of tuition fees on transition from high school to university in germany. *Economics of Education Review*, 41:14 – 23.

Buckles, K., Hagemann, A., Malamud, O., Morrill, M., and Wozniak, A. (2016). The effect of college education on mortality. *Journal of Health Economics*, 50:99 – 114.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, 62(12):3439–3449.

Buser, T. and Yuan, H. (2016). Do Women give up Competing more easily? Evidence from the Lab and the Dutch Math Olympiad. Discussion Paper TI 2016-096/I, Tinbergen Institute.

Butler, J. V. (2016). Inequality and relative ability beliefs. *The Economic Journal*, 126(593):907–948.

Caliendo, M., Cobb-Clark, D. A., and Uhlendorff, A. (2015). Locus of control and job search strategies. *Review of Economics and Statistics*, 97(1):88–103.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, 89(1):306–318.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cameron, C., Gelbach, J., and Miller, D. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882.

Casas-Arce, P. and Martínez-Jerez, F. A. (2009). Relative performance compensation, contests, and dynamic incentives. *Management Science*, 55(8):1306–1320.

Cebi, M. (2007). Locus of control and human capital investment revisited. *The Journal of Human Resources*, 42(4):919–932.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660.

Compte, O. and Postlewaite, A. (2004). Confidence-enhanced performance. *American Economic Review*, 94(5):1536–1557.

Costa-Gomes, M. A., Huck, S., and Weizsaecker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, 88(0):298–309.

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

Dahl, D. W. and Smimou, K. (2011). Does motivation matter?: On the relationship between perceived quality of teaching and students' motivational orientations. *Managerial Finance*, 37(7):582–609.

Damgaard, M. T. and Nielsen, H. S. (2017). The use of nudges and other behavioural approaches in education. EENEE Analytical Report 29, Prepared for the European Commission.

Davies, R., Heinesen, E., and Holm, A. (2002). The relative risk aversion hypothesis of educational choice. *Journal of population economics*, 15(4):683–713.

Dawson, C. (2017). The upside of pessimism: Biased beliefs and the paradox of the contented female worker. *Journal of Economic Behavior & Organization*, 135:215 – 228.

Dechenaux, E., Kovenock, D., and Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, 18(4):609–669.

Dee, T. S. (2014). Stereotype threat and the student-athlete. *Economic Inquiry*, 52(1):173–182.

DesJardins, S. L. and Toutkoushian, R. K. (2005). Are students really rational? the development of rational thought and its application to student choice. In Smart, J. C., editor, *Higher Education: Handbook of Theory and Research*, pages 191–240. Springer Netherlands, Dordrecht.

Dohmen, T. and Falk, A. (2010). You get what you pay for: Incentives and selection in the education system. *The Economic Journal*, 120(546):F256–F271.

Dohmen, T. and Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, 101(2):556–90.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review*, 101(5):1739–74.

Dulleck, U. and Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1):5–42.

Dwenger, N., Storck, J., and Wrohlich, K. (2012). Do tuition fees affect the mobility of university applicants? evidence from a natural experiment. *Economics of Education Review*, 31(1):155 – 167.

Eccles, J. S. and Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1):109–132.

Ederer, F. (2010). Feedback and motivation in dynamic tournaments. *Journal of Economics & Management Strategy*, 19(3):733–769.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–138.

Elsner, B. and Isphording, I. (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics*, 35(3).

Eriksson, T., Poulsen, A., and Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16(6):679–688.

Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3):532 – 545.

Falk, A. and Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57.

Fang, H. and Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In Benhabib, J., Jackson, M., and Bisin, A., editors, *Handbook of Social Economics*, volume 1A, pages 133–200. North Holland, The Netherlands.

Fedor, D. B., Davis, W. D., Maslyn, J. M., and Mathieson, K. (2001). Performance improvement efforts in response to negative feedback: The roles of source power and recipient self-esteem. *Journal of Management*, 27(1):79–97.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2):117–140.

Filippin, A. and Paccagnella, M. (2012). Family background, self-confidence and economic outcomes. *Economics of Education Review*, 31(5):824–834.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Ford, J. B., Joseph, M., and Joseph, B. (1999). Importance-performance analysis as a strategic tool for service marketers: the case of service quality perceptions of business students in new zealand and the usa. *Journal of Services Marketing*, 13(2):171–186.

Fryer, R., Levitt, S., List, J., and Sadoff, S. (2012). Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. Working Paper 18237, National Bureau of Economic Research.

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

Fryer, R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *Quarterly Journal of Economics*, 129(3):1355 – 1407.

Fryer, R. G., Levitt, S. D., and List, J. A. (2008). Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *American Economic Review*, 98(2):370–75.

German Research Foundation (2016a). Excellence initiative (2005-2017) - general information. `http://www.dfg.de/en/research_funding/programmes/excellence_initiative/general_information/index.html`. Accessed: April 12, 2016.

German Research Foundation (2016b). Institutional strategies (2005-2017). `http://www.dfg.de/en/research_funding/programmes/excellence_initiative/institutional_strategies/index.html`. Accessed: April 12, 2016.

Gershkov, A. and Perry, M. (2009). Tournaments with midterm reviews. *Games and Economic Behavior*, 66(1):162–190.

Gibbons, S., Neumayer, E., and Perkins, R. (2015). Student satisfaction, league tables and university applications: Evidence from britain. *Economics of Education Review*, 48:148 – 164.

Gill, D., Prowse, V., Kissova, Z., and Lee, J. (2016). First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. Discussion Paper 783, Oxford Department of Economics.

Greiner, B. (2004). An online recruitment system for economic experiments. In Kremer, K. and Macho, V., editors, *Forschung und wissenschaftliches Rechnen*, pages 79–93. Ges. fuer Wiss. Datenverarbeitung, Goettingen.

Griffith, A. and Rask, K. (2007). The influence of the us news and world report collegiate rankings on the matriculation decision of high-ability students: 1995–2004. *Economics of Education Review*, 26(2):244 – 255.

Grossman, Z. and Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, 84(2):510–524.

Gürtler, O. and Harbring, C. (2010). Feedback in tournaments under commitment problems: Experimental evidence. *Journal of Economics & Management Strategy*, 19(3):771–810.

Guyon, N., Maurin, E., and McNally, S. (2012). The effect of tracking students by ability into different schools: A natural experiment. *Journal of Human Resources*, 47(3):684–721.

Hannan, L., Krishnan, R., and Newman, A. (2008). The effects of disseminating relative performance feedback in tournament and individual performance compensation plans. *The Accounting Review*, 83(4):893–913.

Hannan, L., McPhee, G., Newman, A., and Tafkov, I. (2013). The effect of relative performance information on performance and effort allocation in a multi-task environment. *The Accounting Review*, 88(2):553–575.

Hanushek, E. and Rivkin, S. (2006). Teacher Quality. In Hanushek, E. and Welch, F., editors, *Handbook of the Economics of Education*, volume 2, pages 1051–1078. Elsevier.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 14(3):351–388.

Hanushek, E. A. and Wößmann, L. (2006). Does educational tracking affect performance and inequality? differences- in-differences evidence across countries. *The Economic Journal*, 116(510):C63–C76.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1):81–112.

Hüber, F. and Kübler, D. (2011). Hochschulzulassungen in deutschland: Wem hilft die reform durch das „dialogorientierte serviceverfahren“? *Perspektiven der Wirtschaftspolitik*, 12(4):430–444.

Hübner, M. (2012). Do tuition fees affect enrollment behavior? evidence from a ‘natural experiment’ in germany. *Economics of Education Review*, 31(6):949 – 960.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4):451 – 464. European Association of Labour Economists 23rd annual conference, Paphos, Cyprus, 22-24th September 2011.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3):411–482.

Heineck, G. and Anger, S. (2010). The returns to cognitive abilities and personality traits in germany. *Labour Economics*, 17(3):535–546.

Hoelzl, E. and Rustichini, A. (2005). Overconfident: Do you put your money on it? *The Economic Journal*, 115(503):305–318.

Hogarth, R. M. and Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1 – 55.

Horstschräer, J. (2012). University rankings in action? the importance of rankings and an excellence competition for university choice of high-ability students. *Economics of Education Review*, 31(6):1162 – 1176.

Hossler, D., Braxton, J., and Coopersmith, G. (1989). Understanding student college choice. In Smart, J., editor, *Higher Education: Handbook of Theory and Research*, volume 4, pages 231–288. Kluwer, The Netherlands.

Hoxby, C. M. (2009). The changing selectivity of american colleges. *Journal of Economic Perspectives*, 23(4):95–118.

Hoxby, C. M. and Terry, B. (1999). Explaining Rising Income and wage Inequality Among the College Educated. NBER Working Papers 6873, National Bureau of Economic Research, Inc.

Ilgen, D. R., Fisher, C. D., and Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64:349–371.

Jalava, N., Joensen, J. S., and Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196.

Kerr, S. P., Pekkarinen, T., and Uusitalo, R. (2013). School tracking and development of cognitive skills. *Journal of Labor Economics*, 31(3):577–602.

Kluger, A. N. and DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science*, 7(3):67–72.

Koch, A., Nafziger, J., and Nielsen, H. S. (2015). Behavioral economics of education. *Journal of Economic Behavior & Organization*, 115:3–17.

Kosse, F., Deckers, T., Schildberg-Hörisch, H., and Falk, A. (2016). The formation of prosociality: Causal evidence on the role of social environment. HCEO Working Paper 2016-011.

Kremer, M., Miguel, E., and Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, 91(3):437–456.

Kräkel, M. (2008). Emotions in tournaments. *Journal of Economic Behavior & Organization*, 67:204–214.

Krueger, J. and Mueller, R. A. (2002). Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2):180–188.

Kuhnen, C. M. and Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science*, 58(1):94–113.

Kuziemko, I., Buell, R. W., Reich, T., and Norton, M. I. (2014). "last-place aversion": Evidence and redistributive implications. *The Quarterly Journal of Economics*, 129(1):105.

Lam, S. S. and Schaubroeck, J. (2000). The role of locus of control in reactions to being promoted and to being passed over: A quasi experiment. *Academy of Management Journal*, 43(1):66–78.

Lane, A. M., Whyte, G. P., Terry, P. C., and Nevill, A. M. (2005). Mood, self-set goals and examination performance: the moderating effect of depressed mood. *Personality and Individual Differences*, 39:143–153.

Lavecchia, A., Liu, H., and the Handbook of Economics of Education vol. 5., P. O. (2016). Behavioral economics of education: Progress and possibilities. In Hanushek, E., Machin, S., and Woessmann, L., editors, *Handbook of Economics of Education, Vol. 5*, pages 1–74. North-Holland, The Netherlands.

Lazear, E. (1977). Education: Consumption or production? *Journal of Political Economy*, 85(3):569–597.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

Lempert, K. M. and Phelps, E. A. (2014). Chapter 12 - neuroeconomics of emotion and decision making. In Glimcher, P. W. and Fehr, E., editors, *Neuroeconomics (Second Edition)*, pages 219 – 236. Academic Press, San Diego, second edition edition.

Levitt, S., List, J., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *The American Economic Review*, 90(2):426–432.

Malamud, O. and Pop-Eleches, C. (2011). School tracking and access to higher education among disadvantaged groups. *Journal of Public Economics*, 95(11):1538 – 1549. Special Issue: International Seminar for Public Economics on Normative Tax Theory.

Malmendier, U. and Tate, G. (2005). Ceo overconfidence and corporate investment. *The Journal of Finance*, 60(6):2661–2700.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3):280–295.

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., and Peschar, J. L. (2006). Oecd's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6(4):311–360.

Mas, A. and Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1):112–45.

Milligan, K., Moretti, E., and Oreopoulos, P. (2004). Does education improve citizenship? evidence from the united states and the united kingdom. *Journal of Public Economics*, 88(9 - 10):1667 – 1695.

Mobius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2011). Managing self-confidence: Theory and experimental evidence. *NBER Working Papers*, 17014.

Mueller, R. E. and Rockerbie, D. (2005). Determining demand for university education in ontario by type of student. *Economics of Education Review*, 24(4):469 – 483.

Multrus, F. (2004). *Fachkulturen: Begriffsbestimmung, Herleitung und Analysen*. PhD thesis, University of Konstanz.

Murphy, K. R. and Cleveland, J. N. (1995). *Understanding Performance Appraisal*. Sage, Thousand Oaks.

Murphy, R. and Weinhardt, F. (2014). Top of the class: The importance of ordinal rank. CESifo Working Paper Series 4815, CESifo Group Munich.

Newey, W. K. (1987). Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics*, 36(3):231–250.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Niederle, M. and Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2):129–44.

Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85(5):678–707.

Noftle, E. E. and Robins, R. W. (2007). Personality predictors of academic outcomes: big five correlates of gpa and sat scores. *Journal of personality and social psychology*, 93(1):116.

Obermeit, K. (2012). Students' choice of universities in germany: structure, factors and information sources used. *Journal of Marketing for Higher Education*, 22(2):206–230.

Oreopoulos, P. and Salvanes, K. (2011). Priceless: The nonpecuniary benefits of schooling. *The Journal of Economic Perspectives*, 25(1):159–184.

O'Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C., and Popović, Z. (2014). Brain points: A growth mindset incentive structure boosts persistence in an educational game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3339–3348. ACM.

Park, Y. and Santos-Pinto, L. (2010). Overconfidence in tournaments: evidence from the field. *Theory and Decision*, 69(1):143–166.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., and Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, 26(6):784–793.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661.

Prendergast, C. J. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37:7–63.

Reuben, E., Wiswall, M., and Zafar, B. (2017). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal*, pages n/a–n/a.

Rivkin, S., Hanushek, E., and Kain, J. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458.

Rosenberg, M. (1965). *Society and the adolescent self-image. Princeton, NJ: Princeton University Press.* Princeton University Press, Princeton, NJ.

Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied*, 80(1):1.

Santos-Pinto, L. (2008). Positive self-image and incentives in organisations. *The Economic Journal*, 118(531):1315–1332.

Schaafsma, J. (1976). The consumption and investment aspects of the demand for education. *The Journal of Human Resources*, 11(2):233–242.

Schneider, D. J. (2004). *The Psychology of Stereotyping.* Guilford Press, New York.

Schwardmann, P. and Van der Weele, J. J. (2016). Deception and self-deception. *Tinbergen Institute Discussion Papers*, 2016-12.

Silles, M. A. (2009). The causal effect of education on health: Evidence from the united kingdom. *Economics of Education Review*, 28(1):122 – 128.

Simeaner, H., Ramm, M., and Kolbert-Ramm, C. (2013). Datenalmanach studierendensurvey 1993 - 2013. studiensituation und studierende an universitäten und fachhochschulen. Technical report, Arbeitsgruppe Hochschulforschung, Universität Konstanz.

Spiess, C. K. and Wrohlich, K. (2010). Does distance determine who attends a university in germany? *Economics of Education Review*, 29(3):470 – 479.

Spinnewijn, J. (2015). Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association*, 13(1):130–167.

Steel, P. (2007). The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1):65–94.

Tafkov, I. D. (2013). Private and public relative performance information under different compensation contracts. *The Accounting Review*, 88(1):327–350.

Teisl, M. F., Rubin, J., and Noblet, C. L. (2008). Non-dirty dancing? interactions between eco-labels and consumers. *Journal of Economic Psychology*, 29(2):140 – 159.

The Economist (2015). Top of the class. competition among universities has become intense and international, march 28.

Todd, P. E. and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113:F3–F33.

Tran, A. and Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, 96(9-10):645–650.

Trautwein, U., Lüdtke, O., Köller, O., and Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: how the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, 90(2):334–349.

von Collani, G. and Herzberg, P. Y. (2003). Eine revidierte fassung der deutschsprachigen skala zum selbstwetgefühl von rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie*.

Wagner, V. (2016). Seeking Risk or Answering Smart? Framing in Elementary Schools. Discussion Paper 227, Düsseldorf Institute for Competition Economics (DICE).

Wagner, V. and Riener, G. (2015). Peers or Parents? On Non-Monetary Incentives in Schools. DICE Discussion Papers 203, Heinrich-Heine-Universität Düsseldorf, Düsseldorf Institute for Competition Economics (DICE).

Weiler, W. C. (1996). Factors influencing the matriculation choices of high ability students. *Economics of Education Review*, 15(1):23 – 36.

Wilhelm, O., Hildebrandt, A. H., and Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, 4(433).

Wiswall, M. and Zafar, B. (2015). How do college students respond to public information about earnings? *Journal of Human Capital*, 9(2):117–169.

Wolinsky, A. (1995). Competition in markets for credence goods. *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft*, 151(1):117–131.

Yeager, D. S., Johnson, R., Spitzer, B. J., Trzesniewski, K. H., Powers, J., and Dweck, C. S. (2014). The far-reaching effects of believing people can change: Implicit theories of personality shape stress, health, and achievement during adolescence. *Journal of Personality and Social Psychology*, 106(6):867 – 884.