UNIVERSITÄT
BERN

Three essays in behavioral and managerial economics

Inauguraldissertation zur Erlangung der Würde eines

Doctor rerum oeconomicarum

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Universität Bern

vorgelegt von

Andrea Essl

Bern, September 2015

Die Fakultät hat diese Arbeit am 25.02.2016 auf Antrag der Gutachter Prof. Dr. Frauke von Bieberstein, Prof. Dr. Florian Englmaier und Prof. Dr. Eberhard Feess als Dissertation angenommen, ohne damit zu den darin ausgesprochenen Auffassungen Stellung nehmen zu wollen.

Acknowledgements

I would like to express my gratitude to all who supported me in the process of completing this dissertation project.

First and foremost, I would like to thank my first supervisor Prof. Dr. Frauke von Bieberstein. Frauke has supported me not only by providing funding for all my projects, but also academically and emotionally through the demanding challenge of finishing this thesis. I would like to express my very sincere gratitude for her continuous support of my research, for her motivation, guidance and encouragement. I further thank her for the inspiring and helpful discussions and comments on my research projects. Frauke gave me the moral support and freedom I needed to move on during the most difficult time when working on this thesis. I would also like to thank her for giving me the opportunity and financial support to visit the University of Chicago, various seminar, and workshops.

In addition, I am extremely grateful to Prof. Dr. Florian Englmaier for being my second supervisor and for his valuable time. I would also like to express my gratitude to Prof. Dr. Michael Kosfeld for being a co-author, for his helpful comments and discussions as well as his encouragement.

The years of my study in Bern would have been tedious and less enriching if there were not my colleagues at the University of Bern. I am especially grateful to Stefanie Jaussi, Dr. Nils Kraiczy, Dr. Michael Dorn, Dr. Martin Stähle, Oliver Pahud, Sonja Erni, Maruscha Heyner, Lucian Hunger, Simon Roth, Katja Rubin and Markus Kröll for the stimulating discussions, helpful suggestions, editing advice, and for all the great time we have had in the last four years.

I would also like to thank my friends and roommates for listening, offering me advice and supporting me through this entire process. Special thanks to: Isabella, Maria, Nils, Martin, Irmgard and Laura. Thank you for being there whenever I needed a friend.

Above all, I would like to extend my gratitude to my beloved boyfriend, Sébastien, my parents, Renate and Georg, my brother, Christian, and my nephew, Paul, for their unconditional love and continuous support. They have always been there for me and believed in me. I dedicate this thesis to them.

Bern, September 2015
Andrea Essl

Contents

List of Figures

List of Tables

1 Executive Summary

Behavioral economics combines findings and methods from both psychology and economics to empirically test and modify traditional economic theory. This discipline has identified a variety of biases in human behavior, which are now recognized to be important source of errors in managerial decisions. According to the evidence from economic experiments, humans depart from concepts postulated by standard economic theory such as self-interest and unbounded rationality. Contrarily, results show that people often have to trade with behavioral biases and further try not only to maximize their own outcome but are also concerned about the welfare of others. This observed human behavior has important implications for the understanding of organizations. This thesis is a collection of three essays on behavioral economics using experimentation to analyze the role of pro-sociality, incentive frames, and non-monetary incentives in organizations. The three essays are related to one another by the applied method of an experimental approach. Further, all the presented results deviate from the standard economic predictions and contribute to the improvement of managerial decisions in organizations.

Essay 1, which is a joint work with Frauke von Bieberstein, Michael Kosfeld, and Markus Kröll, studies reciprocal behavior and its implications for sales performance. In particular, the essay examines whether a salesperson's willingness to reciprocate is related to his or her sales performance in real-world sales interactions. This question is analyzed by combining individual behavior in the trust game with a unique field data set on the individual sales performance of the same individuals. Salespeople from a large Austrian retail chain participated in an anonymous, one-shot trust game. This experimental game allows the measurement of an individual's inclination towards reciprocity in a controlled environment. Results show that reciprocal salespeople sell more per customer. A more detailed analysis suggests that this effect is driven by salespeople with an inclination towards reciprocal fairness who are mainly responsible for high-consulting-intensive products. The intuition behind these results is the following: given that reciprocal salespeople try to truly understand the customer's needs to give valuable advice, they might be more able to convey that they are trustworthy and cross-sell related products, resulting in higher revenues per sale. In contrast, salespeople, who are more likely to refrain from opportunistic behavior, might not consider to identify the customer's requirements as necessary and in turn customers do not trust them and buy less. Furthermore, the findings reveal that hyper-reciprocal individuals generate, on average, the lowest revenues per sale. Hyper-reciprocal salespeople may be more concerned about their customers' needs than about their own sales targets and the company's interest. These results emphasize that good salespersons should find the balance between caring about a client's needs and his or her own sales success. Given the positive relationship between reciprocity and revenues per sale, the essay further investigates the question of whether the retail chain benefits from employing reciprocal salespeople. The results demonstrate that in the area where consulting really matters, reciprocal salespeople have a fewer number of sales and that this negative effect of reciprocity on the number of sales outweighs the positive effect on revenues per sale, leading to lower total revenues. One explanation for this result is that in order to give valuable advice, reciprocal salespeople might spend more time with each customer, implying that they have less time for other customers. The findings of this essay do not only show that reciprocal behavior observed in the trust game has predictive power for sales performance in the field but also allow to derive recommendations for practitioners. For example, managers can use experimental games to learn more about the personal characteristics of their workforce which in turn allows them to optimize task assignment, employee training, and the recruiting process.

Contrary to Essay 1, which examines the role of reciprocal behavior in retail, Essays 2 and 3 investigate when, if, and how different incentives schemes affect motivation and performance. Essay 2, which is a joint work with Stefanie Jaussi, contributes to the literature that shows that seemingly irrelevant factors, such as the way the incentives are framed, can influence an individual's behavior. Further, Essay 2 extends previous research by considering that many economically relevant activities are subject to deadlines, which often put people under time pressure. Additionally, it also contributes to recent evidence demonstrating the importance of individual characteristics. In particular, Essay 2 studies whether and how an individual's degree of loss aversion affects performance under deadline-dependent incentives. These questions are analyzed by means of a laboratory experiment in which participants had to work on a real effort task under two payoff-equivalent contracts, framed in bonus and malus terms. Depending on the contract, participants received either a bonus when finishing the task within 10 seconds or a malus when they were unable to complete the task within the 10-second deadline. Incentives were payoff-equivalent across treatments. To identify the predictive power of an individual's degree of loss aversion, a loss aversion test was implemented. The experimental findings suggest that participants with a high level of loss aversion working under a deadline-dependent malus incentive scheme perform worse than all other individuals. The reason behind this result is that individuals with a high level of loss aversion have a higher incentive to avoid a malus than individuals with a low level of loss aversion. Thus, they are more prone to choking under pressure. In addition, the findings reveal that performance differences are driven by both the response time and correct answers. Notably, individuals with a high level of loss aversion gave less correct answers and needed more time to reply, leading to more accumulated maluses. The results of this experiment highlight the importance of considering individual characteristics, such as loss aversion, when writing contracts in time-pressured environments.

Finally, Essay 3 studies gender effects in reactions towards relative symbolic recognition. In particular, this essay examines the questions of whether symbolic recognition in one task influences the subsequent performance in an unrelated task, how the effect differs with positive, negative or no symbolic recognition, and how the reactions towards recognition alter by gender. In an artefactual field experiment, secondary school students had to work on two different tasks. In the first task they had to estimate the number of peas in a bowl, and in the second task, the students had to cut out flyers which promoted an upcoming concert of a university student orchestra. In the experimental treatment, the students received unannounced symbolic recognition after the estimation task. In detail, the top third was honored with a smiley-sticker, the bottom third got a frowny-sticker and the intermediate third did not receive any symbolic recognition but the written information that they performed averagely. Students in the control treatment received neither symbolic recognition nor performance feedback after the estimation task. The essay demonstrates that the response to different symbolic recognition types is heterogeneous across genders. Compared to the students in the control group, the female non-recipients as well as the females who were rewarded by a smiley-sticker in the estimation task significantly improved their performance in the following flyer-cutting task. Moreover, results show that there is no spillover effect of the different recognition types on males' performance. The experimental findings suggest that gender differences in the reaction towards relative symbolic recognition depend on whether the task is more female- or more male-oriented. Hence, when designing incentives the consideration of gender-specific preferences can contribute to the reduction of the still existing gender gap.

2 Essay 1: Does reciprocity sell? Evidence from a lab-in-the-field experiment

Andrea Essl, Frauke von Bieberstein, Michael Kosfeld, Markus Kröll *

Abstract

This paper examines the relationship between an individual's willingness to reciprocate and his or her sales performance in real-world sales interactions. By combining experimental measurements from a trust game with field data, we show that reciprocal salespeople sell more than opportunistic salespeople per customer when consulting matters. However, our data indicate that hyper-reciprocal salespeople, who ended up with less money than their counterpart in the trust game, generate relatively low net-revenues per sale. This highlights the importance of finding a balance between caring about a customer's needs and individual sales targets. Moreover, we find that reciprocity is negatively related to the number of sales, most likely because deeply understanding the customer's needs takes up a lot of time. Thus, fewer customers can be served. When selling high-consulting-intensive products, the negative effect of reciprocity on the number of sales even outweighs the positive effect on net-revenues per sale, resulting in marginally significantly lower total net-revenues. Besides providing a deeper insight into the relationship between reciprocity and sales success, this paper further contributes to the debate on the external validity of laboratory experiments by showing that reciprocal behavior in the trust game has explanatory power in non-laboratory environments.

Keywords: reciprocity, sales performance, trust game, lab-in-the-field experiment, behavioral economics

*Andrea Essl (andrea.essl@iop.unibe.ch) and Frauke von Bieberstein (frauke.vonbieberstein@iop.unibe.ch) are from the Institute of Organization and Human Resource Management, University of Bern. Michael Kosfeld (kosfeld@econ.uni-frankfurt.de) and Markus Kröll (kroell@econ.uni-frankfurt.de) are from the Chair of Organization and Management, Goethe University of Frankfurt. We would like to thank all salespeople who participated in our research and the management who supported the project. We are grateful to Reto Obermatt, Iwan Barankay, Uri Gneezy, Steven Levitt and participants of the internal seminar at the University of Bern, the CUSO workshop, and the Brown Bag Lunch at the Goethe University of Frankfurt for helpful comments. All errors are ours.

2.1 Introduction

Who is selling more per customer: Opportunistic salespeople who only think about their own sales targets or reciprocal salespeople who also have their customers' interest in mind? And does a positive effect on revenues per sale translate into higher total revenues? In this paper we combine experimental measurements from a trust game with a unique data set of individual sales performance to investigate these issues.

The topic is important for several reasons. First, all economic activity ultimately depends on the exchange of goods and services. If customers are not willing to buy, no transactions will take place. Understanding the factors that influence sales outcomes is therefore important from a general economic perspective. Second, empirical studies show that social capital like trust and reciprocity are positively correlated with macroeconomic outcomes (e.g., Putnam et al. 1994, Knack and Keefer 1997, Fukuyama 2001, Carter and Castillo 2002, Guiso et al. 2004, Dearmon and Grier 2009, Dincer and Uslaner 2010). However, research on the relationship between reciprocity and individual work performance is still limited (e.g, Barr and Serneels 2009). This paper is the first that links reciprocity, measured by means of a trust game, to individual sales performance in the field. Third, the topic is of high relevance to practitioners. Starting with Saxe and Weitz (1982), the management literature has a long history of examining the optimal level of customer orientation for salespeople. Whereas in the past more customer orientation has been considered better, more recent evidence shows a mixed effect (Franke and Park 2006) and potential problems of customer orientation (Homburg et al. 2011). While these studies use salespersons' multi-item self-reports of customer orientation, we consider the underlying trait of reciprocity measured by behavior in the trust game.

Why should reciprocity matter in sales? In a typical sales transaction, the customer is less informed than the seller (Akerlof 1982). Thus, customers have to trust in a salesperson's advice. A salesperson that reciprocates the customer's trust will try to truly understand the customer's needs to give valuable advice. In contrast, an opportunistic salesperson will try to sell whatever maximizes his or her own income. A typical sales process follows the five stages of need identification, presentation of the product or service, dealing with objections, negotiation, and closing (Jobber and Lancaster 2006). In any of these stages, a reciprocal salesperson might be more attentive to the specific customer's needs. For instance, during need identification the reciprocal salesperson might actively involve the customer, listen attentively, and ask specific questions on requirements. In contrast, an opportunist who already knows the product he or she likes to offer might try to steer the discussion quickly into a predetermined direction.

In this study, we combine individual behavior in the trust game with real-world sales performance data. Salespeople from an Austrian retail chain participated in an anonymous, one-shot trust game introduced by Berg et al. (1995). In this sequential two-player game, the sender can pass money from his or her initial endowment to the responder. This investment is tripled by the experimenter and handed to the responder. Then, the responder can return any amount of money between zero and the tripled amount plus his or her initial endowment to the sender. Whereas an opportunist returns less money than the amount sent to them by the sender, a reciprocal person returns more. In order to stay as close as possible to the sales process, salespersons' counterparts in the game were Austrians not working for the retail chain. In line with previous research, we define reciprocity as a characteristic of the responder in the trust game (the sales representative) that enhances confidence in the sender (the customer) (Mayer et al. 1995, Ben-Ner and Putterman 2001, Hardin 2002, Caldwell and Clapham 2003).

The retail chain provided us with a unique data set comprising individual sales performance records for each single sale made by each employee over the period March 2012 to March 2014. This rich data set allows us to control for store, month, weekday, and promotion day fixed effects in our regression analysis. In addition, for each salesperson, we received information on tenure, work intensity, the consulting-intensive area he or she was assigned to, and gender. Furthermore, together with the trust game we conducted a survey covering different aspects that have been shown to influence sales performance like questions on education, body height, the Big 5 personality dimensions, and risk and time preferences.

Most empirical studies use surveys to gather information on individuals' social preferences. However, survey-based measurements of pro-sociality may evaluate stated rather than actual behavior and traits (Fehr et al. 2003). Furthermore, in the field, several factors, such as reputation, competition, and information constraints, can lead to confounded measurements of pro-social behavior (Fehr and Leibbrandt 2011). Laboratory measurements allow these issues to be overcome and provide an alternative convenient way for eliciting pro-social behavior. The key advantages of laboratory experiments are the controlled environment, their replicability, and the fact that participants' decisions have monetary consequences that counterbalance the social desirability bias (Fehr et al. 2003, Fehr and Leibbrandt 2011).

Given our research question, our main dependent variable is net-revenues per sale, i.e., revenues net of refunds. Building on prior research of the effect of reciprocity on individual work performance (Barr and Serneels 2009) and on the management literature of customer orientation (Franke and Park 2006, Homburg et al. 2011), we predict a positive effect of reciprocity on net-revenues per sale. This effect should be particularly present for products classified as high-consulting-intensive by the firm. These products are often more expensive, not self-explanatory, and related to more far-reaching decisions than products that require less advice. Thus, when buying a high-consulting-intensive product, customers need thorough personal advice. Therefore, we predict that especially in areas where customers require guidance, reciprocal salespeople are able to sell more per customer. In addition, we consider the effect of reciprocity on total net-revenues. This effect will be determined by the interplay between net-revenues per sale and the number of sales. Regarding the number of sales, we see competing predictions. On the one hand, customers might sense the reciprocity of the salesperson and be more willing to buy from a reciprocal salesperson. On the other hand, giving detailed advice is time-consuming and thus might result in a lower number of sales for reciprocal salespeople.

Our main findings can be summarized as follows. We find that our experimental measurement of reciprocity is an important predictor of sales performance in the field. In particular, we show that salespersons with a higher inclination toward reciprocal fairness sell more per customer. A detailed analysis reveals that this effect is driven by reciprocal and particularly equality-minded salespersons responsible for high-consulting-intensive products. We further observe that hyper-reciprocal salespeople, who returned more than the amount that would lead to an equal split in the trust game, generate, on average, the lowest net-revenues per sale. This effect has also been described in theory where a firm prefers to hire an opportunistic salesperson because a reciprocal salesperson might care too much for the customer and too little for the firm and his or her own sales commission (Lammers 2010). When analyzing the question of whether reciprocity is profitable for the company, we find in a first step that the association between reciprocity and the number of sales is negative and in a second step, that this negative effect of reciprocity on the number of sales even outweighs the positive effect of an individual's willingness to reciprocate on net-revenues per sale, leading to marginally significantly lower total net-revenues in the high-consulting-intensive area.

The positive effect of reciprocity on net-revenues per sale could be due to several reasons. One explanation is that customers somehow sense the opportunistic behavior of the salesperson and buy less. However, if customers distrust the advice, we would expect them not to buy at all rather than to buy a cheaper product. In contrast, we find the opposite result that opportunistic salespeople have more sales closures. Another reason could be that opportunists in general offer cheaper products in the hope of closing the sale quickly to move on to the next customer. Given the incentive system in place that requires the sale to exceed a product-group-specific threshold to be eligible for a commission, we do no find this explanation convincing, either. The most plausible interpretation is based on cross-selling: Given that reciprocal salespeople take the time and energy to really understand the customer's needs, they are more able to cross-sell related products that are of interest to the customer. In contrast, opportunists who might know from the start of the sales process which product they are going to offer might have a harder time listening and offering suitable related products. Furthermore, reciprocal salespeople might be able to establish an environment in which customers feel that their needs are taken seriously and in turn customers buy more in the corresponding store. One natural interpretation for the negative effect of reciprocity on the number of sales is that truly understanding the customer's needs takes up valuable time (Homburg et al. 2011). Thus, reciprocal salespeople might spend so much time with each customer that they cannot serve many.

Our findings are of both managerial and scientific relevance. Determining individual characteristics that drive sales performance has obvious relevance in retail and commerce. Based on our results, we conclude that reciprocal and opportunistic salespersons pursue different sales strategies. Whereas one is focusing on an in-depth understanding of the customer's needs, the other is aiming at a quick and efficient closing of the deal. This information can be helpful for managers in task assignment. Furthermore, the study demonstrates that managers can use simple tools to identify personal characteristics that allow organizations to optimize employee training and the hiring process. From a scientific point of view, our research adds to the debate about the external validity of laboratory experiments.[1] We show that laboratory findings has explanatory power in field environments. Accordingly, we make a general methodological point by providing evidence that reciprocity, measured by the trust game, has predictive power for individual sales performance in the field.

Nevertheless, it should be mentioned that we do not attempt to identify the causal relationship between reciprocity and sales performance. Given possible omitted variables and the difficulty of manipulating a personal characteristic, a causal interpretation of our results should be made with caution. Although we cannot rule out these issues, we examine the relationship between reciprocity and sales performance in a robust way by applying a rich set of econometric specifications. To absorb unobserved heterogeneity across stores and time we include store, month, weekday, and promotion day fixed effects in all our specifications. Additionally, we control for the level of consulting intensity of the products the salesperson is responsible for, tenure, work intensity, and gender. Furthermore, we test whether our results are robust to the inclusion of control variables measuring trusting behavior, body height, education, Big 5 personality dimensions, and individual risk and time preferences. These variables may not only drive sales performance, but may also be correlated with an individual's willingness to reciprocate.

[1]For a detailed discussion on the external validity of laboratory experiments see Levitt and List (2007a) and Levitt and List (2007b).

We can show that our main results are robust to the inclusion of these different control variables.

Our paper contributes to different strands of literature. First, there is a large body of laboratory experiments demonstrating that people not only behave opportunistically as predicted by the standard economic models, but also care about the welfare of others (e.g., Fehr and Gächter 2000, Camerer 2003, Dohmen et al. 2009, Cooper and Kagel 2013). The trust game, designed by Berg et al. (1995), has become a well-grounded approach to measure two important elements of social capital: trusting behavior and an individual's willingness to reciprocate (for an overview see Johnson and Mislin 2011). In the last decade, researchers have investigated whether the trust game can predict naturally occurring outcomes (Karlan 2005, Baran et al. 2010, Serra et al. 2011).[2] For example, Karlan (2005) showed that participants who reciprocated in the trust game were more likely to repay their loans. Baran et al. (2010) used the trust game to measure reciprocity among MBA students and found that the behavior in the game predicted the amount donated to their university. Furthermore, Serra et al. (2011) demonstrated that the reciprocal responders in the trust game possess a higher willingness to work in the nonprofit sector.

Second and most closely related to our study, there is limited research which combined laboratory measurements and field performance data to understand the impact of pro-social characteristics on productivity (e.g., Barr and Serneels 2009, Leibbrandt 2012). Bowles et al. (2001) argued that behavioral traits may be even more important for earnings than standard variables like grades and age. Barr and Serneels (2009) provided evidence for the argumentation of Bowles et al. (2001) by showing that reciprocal behavior, measured in a trust game, is positively correlated with both firm productivity and individual earnings. Unfortunately, they only had individual data on wages and not on productivity. Leibbrandt (2012) observed that shrimps sellers, who cooperated in a public goods game were more able to sell shrimps of similar quality for higher prices than selfish ones. Leibbrandt (2012) further showed that more cooperative shrimp sellers reported that they are more able to signal trustworthiness and thus suggested that an

[2] Several scholars examined the effect of different types of pro-social behavior, measured in other laboratory experiments than the trust game, on field outcomes. List (2006) studied sports card dealers' behavior in the laboratory and in the field. His results showed that nonlocal dealers misrepresented quality significantly more often in the field than in the laboratory. Sellers classified as local dealers, however, behaved similarly in both settings. Benz and Meier (2008) examined how people behaved in a donation experiment and found that pro-social behavior in the laboratory was related to the behavior in the field. Boly (2011) showed that individuals reacted to monitoring by increasing the effort level both in the laboratory and the field. Other authors analyzed the relationship between cooperative behavior, measured in a public goods game, and field outcomes (e.g., Rustagi et al. 2010, Carpenter and Seki 2011, Fehr and Leibbrandt 2011). In contrast to these studies, which indicate a good generalizability from the laboratory to the field, Stoop et al. (2012) observed that fishermen were less cooperative in the field than in the lab.

individual's inclination towards reciprocity is one of the most important characteristics of sales success.

Third, our paper is also related to the psychological and management literature examining whether pro-sociality and customer-oriented selling increase sales performance (e.g., Saxe and Weitz 1982, Franke and Park 2006, Homburg et al. 2011, Grant 2013). Saxe and Weitz (1982) argued that selling-oriented salespeople try to make immediate sales, independently of their clients' needs, while customer-oriented sales representatives avoid action that is in conflict with their clients' interests. Furthermore, they suggested that the success of a customer-oriented approach depends on factors such as the consulting intensity of the product or the market environment. Moreover, in a meta-analysis, Franke and Park (2006) demonstrated that customer-orientation does not necessarily lead to higher sales volumes, due to its costs such as the time spent for identifying the customers' needs. According to Homburg et al. (2011), the relationship between sales performance and a salesperson's customer-orientation has an inverted U-shape. In particular, they observed that salespeople, who indicated customer-orientation levels higher than the optimum, served fewer customers than others. Grant (2013) observed that salespeople, who exhibited high degrees of unconditional pro-sociality, generated either the highest or the lowest revenues. He suggested that poorly performing pro-social sales representatives are more focused on their customers' concerns than on their sales targets, while outstanding salespeople are able to find the balance between their customers' needs and their own interests. Thus, it is important to identify salespeople who care equally about their clients' concerns and their own sales success.

The remainder of the paper is structured as follows. Section 2.2 provides a detailed description of the retail chain and the field data. Section 2.3 outlines the experimental design and procedure. Section 2.4 presents the results. Section 2.5 summarizes and concludes.

2.2 Institutional setting and field data

2.2.1 Institutional setting

As of September 2013, the company comprised 66 retail stores in Austria, of which 23 were flagship stores. In contrast to the traditional regular stores, flagship stores have larger surface areas, product ranges and numbers of employees. Out of the 3147 employees, who were directly assigned to one of the stores, 1785 were active in sales. The sales force consists of sales representatives, apprentices, expert advisers, project advisers, team leaders, and auxiliaries. Since this study concentrates on sales representatives, this position is described in more detail. In September 2013, 1318 sales representatives worked for the company. Around 71% of the sales representatives were full-time employees and half of them worked in flagship stores.

The company classifies products as low-, medium- or high-consulting-intensive. In general, more expensive products are grouped in more consulting-intensive areas. Employees can sell from different areas; however, they are specialized in one of the areas. Therefore, sales representatives are assigned to the area from which they sell the most. On average, 73% of sales take place in the appointed consulting-intensive area. While women and men were nearly evenly distributed among all sales representatives, this is not the case for the different consulting-intensive areas. Table 2.1 shows that women were mostly responsible for lower consulting-intensive products (87%), while they were under-represented in the high-consulting-intensive area (16%). With a share of 56% of female sales representatives, gender was nearly evenly distributed in the medium-consulting-intensive area. Sales representatives in the low- and medium-consulting-intensive area had a tenure of 8 years, and those in the high-consulting-intensive area of 5.7 years. On average, females stayed longer in the company (8.8 years) than males (6.2 years).

TABLE 2.1: Sales representatives by consulting intensity

Consulting intensity	Share of sales representatives	Tenure in years	Share of women
All	100%	7.5	52%
High-consulting-intensive	30%	5.7	16%
Medium-consulting-intensive	44%	8.2	56%
Low-consulting-intensive	26%	8.5	87%

To foster personalized interaction with customers, the company implemented a commission system, which is based on the individual's revenues. The incentive system works as follows: each salesperson is encouraged to attach a personalized sticker to any product he or she has sold. Once a sticker is registered at the cash point, all further sales by the same customer are assigned to the sales representative who put the sticker on the product. The criteria for receiving a commission are based on a receipt threshold and a monthly revenue threshold. The product with the highest price on the receipt determines whether the low-, medium- or high-consulting-intensive product area is relevant for the threshold assignment. The required minimum for low-/medium-/high-consulting-intensive products are 40 Euro/60 Euro/180 Euro, respectively. Only when the product with the highest share on the receipt exceeds the receipt threshold, the receipt is qualified for commission.[3] Each month, the sum of all revenues of receipts exceeding the receipt threshold is compared to a monthly revenue threshold that is independent of the consulting intensity. Refunds in the account of a sales representative are deducted from his or her monthly generated revenue.[4] Salespeople only receive a commission if the sum of revenues minus refunds during the month exceeds the monthly revenue threshold. For part-time employees revenues are adjusted to full-time equivalents. Salespeople received, on average, a monthly commission payment of 42.11 Euro, with a standard deviation of 55.52 Euro. This accounts for approximately 3.4% of the monthly income. The company provided us with the individual performance records from this commission system including all sales independent of whether or not they surpassed the threshold levels. In Section 2.2.2, data is described in more detail.

2.2.2 Field data

The company granted us access to a rich data set on daily sales performance data and information on store characteristics. Data on store characteristics comprises the store location, whether the store is a regular or a flagship store, and the number of employees in each store. Data on the individual level includes employees' tenure, job title, gender, and the assigned store. In addition, for the period March 2012 to March 2014, we have got information on company-wide promotion days, on the individual daily working time, and on individual performance data. This data comprises each receipt assigned to an employee's record as described in Section 2.2.1. In particular, we have the following information: revenues per receipt, the number of sales, the product group of the item with the highest share on the receipt and thus whether the product is high-, medium- or low-consulting-intensive, whether the sale was a cash point revenue or a sale from a customer pick-up, the number and volume of refunds, and the number and volume of commissions received per sale conditional on the monthly minimal revenue being met.

In order to be able to compare the performance of part- and full-time sales representatives, we adjusted the data based on the Austrian 40-hour week which equals a full-time equivalent.[5] Our main variable of interest is daily net-revenues per sale. Daily net-revenues comprise cash point revenues plus revenues generated from customer pick-ups

[3] If, for example, three products are in the customer's trolley and the first product, which belongs to the high-consulting-intensity product group, costs 20 Euro, the second one is also from the high-consulting-intensity area and makes 100 Euro, and the third, which is assigned to the medium-consulting-intensity area, has a price of 70 Euro, then the second product from the high-consulting-intensity area forms the main part and thus the receipt does not qualify for commission. Instead, if the price of the third product was 101 Euro, the receipt would count for the commission payment.

[4] According to the refund policy of the retail chain, products that are returned within four weeks of purchase with the original receipt and in their original packaging are refunded with the purchase price.

[5] Since the company provided us with the individual time recordings, we adjusted performance data by multiplying them by the working hours of a full-time equivalent and dividing them by the actual hours worked of the corresponding employee.

minus refunds. Since customer pick-ups are recorded on the day the client picked up the product and not when the sales representative actually makes the transaction, we evenly distribute the generated revenue of the picked-up item over all workdays of the sales representative in the last 30 days. We follow the same procedure for adjusting the daily number of sales. Refunds are also recorded on the day the customer actually returns the product. Thus, we evenly allocate refunds over all workdays in the last 30 days on which the sales representative worked and on which he or she generated revenues higher than or equal to the refund itself. We choose the 30-day threshold as it is consistent with the company's refund policy.[6] Daily average net-revenues per sale are then calculated using the harmonic mean.

2.3 Experimental design

2.3.1 Procedure

Besides providing us with the field data, the company supported us in conducting a lab-in-the-field experiment with its workforce. Because we had to ensure that we did not interfere with the daily business of the company, the study was undertaken via mail correspondence.[7] Owing to this unconventional way of conducting the experiment and the inexperienced participant pool, we provided information brochures one month before the experiment was conducted. These preparatory mailings were aimed at motivating the staff to take part in the paid study.[8]

In November 2013, all employees received the experimental material via mail.[9] The postal item included a cover letter, instructions, decision sheets, a survey, an identification card for payment and a postpaid envelope for return.[10] In the mailing, participants were informed about the rules of the experiment and the payment procedure. We stressed that all personal information is treated in strict confidence and anonymity. For the purpose of avoiding collusions during the decision-making process, mailings were sent to employees' home addresses and it was explicitly underlined in the instructions that consultation with family members, friends or co-workers during the decision-making process would

[6] The results of our analysis do not change when applying 7-, 14- or 40-day thresholds.

[7] Other economic studies that integrated experiments in surveys are those of Fehr et al. (2003), Bellemare and Kröger (2007), and Falk et al. (2013).

[8] As we are interested in exploring how people make economic decisions, we pointed out the importance of autonomous decision-making. In addition, we mentioned that there are no wrong or right answers and that earnings are based on both one's own decisions and the decisions of other participants. Furthermore, we highlighted that participation is voluntary and data privacy is guaranteed. In particular, it was stated that all data is restricted to scientific purposes and not given to any member of the company or any other third party. Moreover, we underlined that data will be stored anonymously.

[9] The company asked us to let all employees participate in the experiment to avoid complaints of inequality. In the following, only decisions of sales representatives are considered.

[10] An English version of the instructions is available in Appendix A.

harm the research project. While compliance with this rule could not be controlled, store managers mentioned in post-experimental interviews that there had been no obvious collaboration between co-workers in the stores. Additionally, we controlled for agreements by asking participants in the survey whether they had completed the decision sheets by themselves or not. Participants were encouraged, first to read the instructions, then to answer the comprehension questions, and afterwards to complete the decision sheets. In the case of open questions we also invited participating employees to send us an email or to call us.

All documents had to be returned within three weeks via a postpaid envelope. In the instructions as well as in the information brochure it was made clear that each fourth participant would receive the payment in cash for one randomly determined game.[11] The draw decided which participant and which game was paid out. Due to payment procedures and anonymity concerns over the experimenters, management, and co-workers, each participant received a card with an identification (ID) number. The ID number was also printed on the corresponding participant's decision sheets. Participants were encouraged to keep their ID card for the payment, which took place in January 2013. The sealed envelopes with the payoffs and feedback information about the actual decision of their counterparts were sent to the respective store managers. Each envelope was stamped with an ID number. In addition, a list with the ID numbers of the participants who were randomly determined for payment was provided to the store managers. In exchange for their ID card, winners could collect the sealed envelope from their store managers.

2.3.2 Experimental measurement

The experimental design is based on the standard trust game introduced by Berg et al. (1995). At the beginning of the game the sender and the responder were both endowed with 18 Euro. The sender had to decide how much money of his or her initial endowment he or she wanted to transfer to the responder. The choice set of the sender was discretized to 0, 6, 12 or 18 Euro. The transfer was tripled by the experimenter and given to the responder. Contingent upon the sender's transfer, the responder could return any amount between zero and his or her total amount available, which corresponded to the initial endowment of 18 Euro plus three times the investment by the sender. For this reason the possible back transfer could be between 0 and 72 Euro. In the trust game, both players are better off if they cooperate; however, the sender has to trust in the responder's willingness to reciprocate and the responder has to repay. Therefore,

[11] We used a set of different games: a competitiveness game, a trust game, and a public goods game. Only the trust game is part of this study.

the investment serves as a measure of trusting behavior[12] and the back transfer as a measure of positive reciprocity. Payoffs of the sender were determined by the endowment of 18 Euro minus the transfer to the responder plus the back transfer of the responder. The responder's payoffs consisted of his or her initial endowment of 18 Euro plus the tripled transfer minus the amount he or she decided to return.

For eliciting reciprocity we implemented the strategy method. Therefore, the responder had to decide how much to return for all four possible transfers of the sender (0, 6, 12, and 18 Euro). The strategy method was chosen for two reasons. First, it allowed us to gather the complete strategy plan of the responder independently from the sender's actual decision. Thus, we also acquired information about less frequently chosen investments. Second, the strategy method allowed for a simultaneous implementation of the sequential game providing a simplification of the experimental procedure.[13] Since we are interested in measuring reciprocity as well as trusting behavior, we asked employees to play both the role of the responder and the sender. To make instructions easier, we decided to frame the decisions as two separate games. In the first game participants played the responder role and in the following game they acted as the sender. Burks et al. (2003) found that playing both roles reduces both investments and returns. However, they also showed that this effect was only present when participants were informed about playing both roles prior to making their decisions. Since our main research question is how an individual's willingness to reciprocate is related to sales performance, we let participants first make their decisions in the role of the responder without mentioning that they also have to play the role of the sender afterwards. Thus, we assume that at least the participants' behavior in the responder role is not very sensitive to playing both roles of the game. However, as the experiment was conducted via mail, we are not able to rule out that the participants did not discover that they have to play sender role after they have made their decision as responder. The study documents also contained two comprehensive questions serving as a means to identify participants who did not understand the instructions.

The trust game allows us to simulate a one-shot sales interaction in the sense that the customer has to trust a sales representative's advice and the salesperson can either take advantage of it or reciprocate. To strengthen this analogy, employees were informed that the other party was a person from Austria who does not work for the retail chain. We

[12]There is an ongoing debate about what the sender's decision actually measures. While Cox (2004) tried to decompose trust from altruism and reciprocity concerns, Karlan (2005) and Schechter (2007) suggested that trust measured in the game is associated with both trust and risk preference.

[13]So far, there has been no clear evidence about whether the strategy method induces different behavior than the direct-response method. While some researchers found that the process of thinking through the behavioral implication of each outcome induces distortions relative to the direct-response method (Güth et al. 2001), others have shown that there is no significant difference between the two methods (Oxoby and McLeish 2004, Brandts and Charness 2011).

elicited the decisions from the employees' counterparts in a separate study conducted with students from the Innsbruck-EconLab. This procedure further counteracted collusive behavior among employees.

Complementary to the experimental measurements, a survey was part of the study. The survey contained, amongst others, items on risk (Dohmen et al. 2011) and time preferences (Vischer et al. 2013), and the 15-items Big 5 inventory (Gerlitz and Schupp 2005). Additionally, we asked for demographic characteristics such as gender, age, education, and body height. The survey not only serves as a back-up and control of the behavioral measures, but also allows us to gather additional information that may influence sales performance.

2.4 Results

This section is structured as follows. After describing the sample, the results of the trust game are presented. Then, we enlist the empirical strategy and analyze how reciprocity is related to sales performance. Finally, we test how robust our results are with respect to different specifications.

2.4.1 Sample characteristics

The study documents were sent to 1369 sales representatives of the Austrian retail chain.[14] Overall, 291 salespeople returned their decision sheets, thus we obtain a response rate of 21.3%. For the sake of measurement accuracy, we drop the observations of nine participants who indicated that they did not complete the decision sheet by themselves, and of 25 participants who did not answer the control questions correctly. The response rate varies across stores, ranging from 5% to 73%. Since it is necessary to control for store fixed effects in our further analysis, we focus on observations of salespeople from stores in which at least two persons returned their decision sheets. This leads to the exclusion of observations from nine salespeople. Accordingly, we include observations from 248 sales representatives in our analysis.

The personnel records provided by the company make it possible to test whether there are self-selection effects regarding individual characteristics and performance. Sales representatives who took part in the study differ from those who did not take part

[14] The number of sales representatives differs from that in Section 2.2.1, because in Section 2.2.1 the analysis is based on individuals who worked as sales representatives in September 2013, while here we include all individuals who obtained the study documents and worked as a sales representative for at least one month between March 2012 and March 2014. For example, several persons had worked as a sales representative first and then were promoted to a expert adviser, project adviser or a team leader.

in several dimensions. Table 2.2 shows the differences between participants and non-participants according to gender, store type, work intensity, tenure, consulting-intensive areas, and performance measurements. The first four columns of Table 2.2 are based on individuals from all stores who received the study documents and worked as a sales representative for at least one month between March 2012 and March 2014, while the last four columns are addressed to a restricted sample based on stores with more than one participant. Irrespectively of whether individuals from stores with fewer than two participants are excluded or not, men, salespeople who worked in flagship stores, and full-time salespeople were significantly less likely to participate than women, salespeople who worked in regular stores, and part-time salespeople, respectively. When we consider salespeople from all stores, those working in the high-consulting-intensive area were significantly less likely to participate, while those from the low-consulting-intensive area were significantly more likely to take part in the study. Interestingly, when taking the restricted sample, the distribution over consulting-intensive areas is not significantly different between participants and non-participants. Table 2.2 further presents sales performance differences between participants and non-participants. In both samples, participants and non-participants do not significantly differ in respect of the mean daily net-revenues per sale, the mean daily net-revenues, and the fraction of days with a sale. However, participants of the study have a significantly higher mean number of sales.[15]

Since we only have data on the degree of reciprocity of sales representatives who returned the documents, it is impossible to rule out selection bias regarding social preferences. Recently, Falk et al. (2013) and Cleave et al. (2013) tested whether there is a self-selection bias in laboratory experiments and came up with the result that students with stronger pro-social preferences are not more likely to participate in experiments than students with weaker pro-social preferences. Following these studies, it seems plausible to assume that also in our sample no self-selection bias based on social preferences occurred. However, we are not able to exclude the bias with the available data.

[15]Moreover, in Appendix B, in Table 2.7 and Table 2.8, we present probit regressions where the dependent variable equals one if the sales representative participated in the study and zero otherwise. Taking the subject pool including individuals from all stores, the estimates reconfirm the descriptives and indicate that there are significant differences between participants and non-participants regarding gender and store type. Considering only those individuals from stores with more than one participant, results show that there are only significant differences between participants and non-participants in respect of the store type. By separately adding performance variables in specifications 2 to 6 of Table 2.7 and Table 2.8, we examine whether the selection bias is caused by different performance measurements. Results reveal that sales representatives who generated higher daily net-revenues were significantly more likely to participate in our study. However, the size effect is nearly equal to zero. This suggests that the bias due to self-selection caused by performance differences is likely to be small. We do not find any differences regarding other sales performance measurements.

Table 2.2: Selection effects

	All stores				Stores with more than one participant			
	All (n=1369)	Non-participants (n=1078)	Participants (n=291)	p-values	All (n=1169)	Non-participants (n=921)	Participants (n=248)	p-values
Gender and job-related characteristics								
Male	48.58%	50.93%	39.86%	.001***	48.25%	50.27%	40.73%	.005***
Flagship store	50.55%	52.69%	42.61%	.002***	51.58%	54.40%	41.13%	.000***
Full-time	71.95%	73.28%	67.01%	.039**	71.09%	72.42%	66.13%	.058*
Tenure in months	89.13	88.40	91.84	.214	89.70	89.12	91.85	.274
	(78.35)	(78.83)	(76.61)		(77.80)	(78.48)	(75.45)	
Consulting intensity								
High	30.09%	31.35%	25.43%	.052*	30.45%	31.38%	27.02%	.214
Medium	43.97%	43.88%	44.33%	.894	43.11%	43.00%	43.55%	.885
Low	25.93%	24.77%	30.24%	.060*	26.09%	25.19%	29.44%	.192
Performance measurements								
Net-revenues per sale	72.49	73.81	67.59	.300	75.15	76.91	68.63	.127
	(112.11)	(125.04)	(34.47)		(120.61)	(134.63)	(34.89)	
Number of sales	12.51	12.34	13.15	.014**	12.37	12.16	13.18	.038***
	(7.25)	7.32	7.01		(7.06)	(7.07)	(6.98)	
Net-revenues	826.54	817.28	860.82	.256	839.69	831.14	871.47	.345
	(502.06)	(493.64)	(531.55)		(505.78)	(496.77)	(537.82)	
Fraction of days with a sale	90.16%	90.28%	89.73%	.374	90.72%	90.90%	90.05%	.370
	(20.07)	(19.42)	(22.37)		(18.99)	(18.10)	(22.05)	

The table reports the shares of non-participants and participants and the means and standard deviations of non-participants and participants. Standard deviations are displayed in parentheses. Two-sided Mann-Whitney tests are used for numerical data and two-sided Fisher's exact tests for categorical data. Significance levels indicate a difference between non-participants and participants. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.4.2 Trust game

We measure an individual's willingness to reciprocate by the average return on investment ratio (RIR). This ratio is defined as the return divided by the investment. Because the strategy method is used for eliciting the behavior in the role of the responder, we obtained three valid RIRs for each participant (for investments of 6, 12, and 18 Euro). Our measurement of reciprocity is the average of these three single RIRs. In cases where investments are zero, no meaningful values can be calculated. These observations are therefore neglected in the average RIR and accounted for separately in the regressions.[16] The main advantage of the RIR is that it is automatically scaled while it controls for differences in the amount available for the back transfer. Table 2.3 shows the summary statistics of the trust game. The average investment is 7.55 Euro, which is 41.94% of the maximal amount available (18 Euro), with a standard deviation of 5.20 Euro. In addition, Table 2.3 indicates that higher investments were reciprocated with higher returns. The mean of the average RIR is 1.55, with a standard deviation of 0.74. This suggests that, on average, responders compensated senders by returning more money than the senders invested. Interestingly, 33.06% of salespeople returned a small amount of money even in the case when the sender transferred nothing. We control for these returns in the following regression analysis, as they can be seen as an indicator for altruistic preferences (Cox 2002, Cox 2004).[17]

TABLE 2.3: Descriptive statistics of the trust game (n=248)

	Mean	S.D.
Investment	7.55	5.20
Return if investment = 0	2.27	3.78
Return if investment = 6	9.94	5.13
Return if investment = 12	18.27	9.12
Return if investment = 18	26.51	13.96
Avg. RIR	1.55	0.74

[16] We also estimate an individual's willingness to reciprocate by running simple OLS regressions of returns on the investments. Results are presented in Section 2.4.6 and in Appendix B.

[17] Empirical evidence shows that individuals drawn from other populations than a students subject pool behave more pro-socially than students (e.g., Fehr and List 2004, Dohmen et al. 2008, Cleave et al. 2013, Falk et al. 2013).

2.4.3 Empirical strategy

Before presenting the results, we discuss important issues regarding our empirical strategy. Many salespeople have working days on which they did not make a single sale because they were involved in other activities like stock management. However, no formal tracking system of these activities exists. Thus, a significant fraction of the performance data has zero value. Due to these circumstances we split our analysis into two parts. First, we use OLS regressions to model outcome variables conditional on nonzero sales. Second, we apply a probit model for estimating the probability of making a sale.[18] In all regression models, robust standard errors are clustered on an individual level. Data are pooled on a daily level. We take the logarithm of net-revenues per sale, the number of sales, and net-revenues because data is very right-skewed and a Box-Cox model gives much larger support for log-linear transformations than for linear transformations. To reduce the error variance and to get meaningful results, we control for month, weekday, promotion day, and store fixed effects in all specifications. In particular, we include month dummies, because data suggests essential seasonal effects. We further add weekday dummies, since data shows that the weekdays Monday, Friday, and Saturday are busier than the other days. Therefore, a sales representative who works either on Monday, Friday or Saturday may perform better than if he or she had worked on another day. The same applies for promotion days, as on these days the number of sales as well as revenues are significantly higher.[19] Moreover, all our specifications include store fixed effects which absorb any variation in outcomes across individuals that are caused by store-level differences such as size, location or product ranges. In addition, since a non-negligible share of participants returned a positive amount even if the sender transferred nothing, we further control for the corresponding back transfer. Moreover, because more expensive products are grouped in more advice-intensive areas, we also control for consulting intensity. Another variable we control for is tenure. Tenure is expected to be important for sales performance, because it is associated with greater expertise and commitment. Finally, we control for gender and whether a salesperson is a full-time or a part-time employee, because the salespeople who took part in the study differ from those who did not participate among these covariates. In conclusion, we present regression models including a rich set of control variables that are determining for sales success and allow the effect of reciprocity to be isolated.

[18] In the econometric literature this approach is also known as the two-part model (Wooldridge 2010, Cameron and Trivedi 2010). The reason why we used this method instead of a Heckman selection model is that zero observations occurred because salespeople did not make a sale even if they had the possibility of doing so. Hence, we have no missing data and thus no sample selection problem. As we are interested in modeling actual, as opposed to potential sales performance, the approach we implemented is more appropriate than a Heckman selection model (Dow and Norton 2003, Madden 2008).

[19] The company offers several promotion days during the year. On these days customers receive a discount on either a product group or on the full range of products.

Besides the non-negligible fraction of days with zero sales and the large error variance, two other important concerns have to be addressed: reverse causality and a possible omitted variable bias. First, it may be that a salesperson's willingness to reciprocate influences sales performance, but it could also be that performance implies reciprocity, leading to reverse causality. However, in line with previous research, we consider an individual's willingness to reciprocate to be a stable characteristic that is not influenced by sales performance (Mayer et al. 1995, Ben-Ner and Putterman 2001, Hardin 2002, Caldwell and Clapham 2003, Carlsson et al. 2014). Moreover, we probe the robustness of our findings for a subsample of the participants, including only salespersons who worked less than one year for the company. This enables us to exclude the possibility that sales performance causes reciprocity, because it seems highly unlikely that an individual's inclination toward reciprocity is already influenced by his or her sales performance after this short period of time. In this analysis, the main effects stay qualitatively robust and hence we suggest that reciprocity triggers sales performance instead of the reverse.[20] Second, a possible omitted variable bias can arise, because unobserved variables may correlate with an individual's willingness to reciprocate and thus drive the estimated effects. One way to address this issue is to use a fixed effects model, which permits the time-invariant component of the error term to be correlated with the regressors. However, by mean-differencing the fixed effects model removes not only unobserved, but also observed time-invariant components. Since we consider reciprocity as a stable characteristic, it is a time-invariant explanatory variable and would be canceled out in a fixed effects model. As our data does not allow an omitted variable bias to be ruled out completely, we test its extent by examining the change in the magnitude of the coefficient of reciprocity in response to the inclusion of further control variables (Altonji et al. 2008, Bellows and Miguel 2009, Kosfeld and Rustagi 2015). In Section 2.4.6, we show that the magnitudes of the reciprocity coefficient are robust and thus that it is likely that the relationship between reciprocity and sales performance is not caused by unobserved variables.

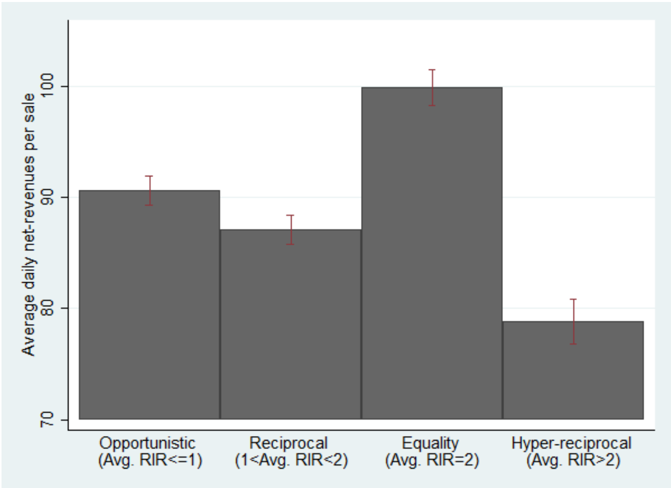
2.4.4 Reciprocity and net-revenues per sale

In this section, we address our main research questions and examine whether there is a positive relation between a salesperson's willingness to reciprocate and revenues per sale in real-world sales interactions, and if so, whether the effect of reciprocity on revenues per sale varies with the consulting intensity of products. The intuition behind the hypothesis that salespeople with a higher inclination toward reciprocity generate more net-revenues per sale is as follows: as reciprocal salespeople may invest more energy to understand the customer's true interests and to give valuable advice, they may be more able to convey

[20]Results available from the authors on request.

that they are trustworthy and in turn cross-sell related products, resulting in higher net-revenues per customer.

Figure 2.1: Net-revenues per sale by behavioral types



The figure shows means and error bars.

For the graphical presentation of our results, we categorize sales representatives according to their average RIR. An average RIR equal to or less than one reveals that the responder is opportunistic and transferred back less than or the same as the amount the sender invested, whereas an average RIR higher than one but less than two indicates that the individual is reciprocal and sent back more than the sender transferred but less than the amount that results in equal final payoffs. An equality-minded person matched his or her final payoffs in a way that both parties ended up with the same amount of money. The corresponding average RIR is equal to two. Finally, a hyper-reciprocal individual with an average RIR higher than two sent back more than the amount that would lead to an equal split and thus ended up with less than the sender. Approximately 30% of the participants behaved in an opportunistic manner, 33% sent back more than they received but less than the equal split, 26% of the participants equalized payoffs, and the remaining 11% behaved hyper-reciprocally and returned more than the equal split. Figure 2.1 shows the mean daily average net-revenues per sale for the different behavioral types. Equality-minded persons sell, on average, 10% to 13% more (100 Euro) than opportunistic and reciprocal salespeople (91 Euro and 87 Euro, respectively). In contrast, hyper-reciprocal individuals have the lowest mean net-revenues per sale (81 Euro). The reason for the low performance of hyper-reciprocal salespeople may be that they care more about others than about themselves. Here, they may care more about the client's needs than about their own commission and the company's interest. This might go so far that they advise customers to go to another store or to buy the item on the Internet.

In Table 2.4, we estimate the effect of reciprocity on net-revenues per sale using pooled OLS regressions. In the regression analysis, we employ the average RIR as a continuous measurement of reciprocity. In all specifications, we control for month, weekday, promotion day, and store fixed effects. In addition, we include the amount returned if the sender transferred 0 Euro, an indicator variable for the consulting-intensive areas, tenure, gender, and a dummy variable, which is one if an employee worked full-time and zero otherwise. Specification 1 shows that the average RIR is significantly and positively correlated with daily average net-revenues per sale. The magnitude of this result is significant as well. Since we take the logarithm of the daily average net-revenues per sale, estimates can be interpreted directly as percentage changes. A shift from a sales representative who retained the entire money (average RIR = 0), independently of the amount he or she received, to a person who sent back the amount he or she received (average RIR = 1), is associated with an approximately 6.1% increase in daily average net-revenues per sale. This means that an equality-driven salesperson (average RIR = 2) sells, on average, approximately 12.2% more per customer than a person who returned nothing (average RIR = 0). In specification 2, when observations from 28 individuals with an average RIR higher than two are excluded,[21] the effect of reciprocal behavior on net-revenues per sale becomes even more pronounced. This confirms the results of the descriptive analysis, which also reveals that net-revenues per sale increase with the degree of reciprocity reaching the peak at an average RIR equal to two.

Next, we investigate the effect of consulting intensity. Consulting intensive areas are represented by an indicator variable with the categories high-, medium-, and low-consulting. The low-consulting-intensive area serves as the reference group. In all four specifications, salespeople who are mainly responsible for medium- and high-consulting-intensive products generate significantly more revenues per sale than those who are assigned to the low-consulting-intensive area. In particular, the effects of being responsible for medium- and high-consulting-intensive products are strong and highly significant. These effects are not surprising, given that customers often require more advice when buying more expensive products. The company has considered this in the formation of the categories: more expensive products are assigned to the areas where consulting matters.

[21] In specifications 2 and 4, for applying store fixed effects, we additionally have to exclude four individuals with an average RIR smaller than or equal to two. This exclusion ensures that there are at least two participants per store.

TABLE 2.4: Log net-revenues per sale - OLS

	1 All	2 Excl. RIR>2	3 All Interactions	4 Excl. RIR>2 Interactions
Avg. RIR	0.061**	0.092**	0.010	-0.025
	(0.028)	(0.043)	(0.044)	(0.069)
Return if investment = 0	0.002	0.002	0.003	0.004
	(0.005)	(0.007)	(0.005)	(0.007)
Medium consulting	0.390***	0.404***	0.319***	0.305**
	(0.048)	(0.052)	(0.112)	(0.152)
High consulting	0.756***	0.804***	0.560***	0.447**
	(0.065)	(0.072)	(0.152)	(0.178)
Full-time	0.130***	0.112**	0.130***	0.103**
	(0.048)	(0.054)	(0.047)	(0.052)
Tenure in months	0.001	0.001	0.001	0.001
	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.193***	0.180***	0.203***	0.198***
	(0.060)	(0.067)	(0.059)	(0.065)
Medium consulting x avg. RIR			0.046	0.081
			(0.059)	(0.096)
High consulting x avg. RIR			0.130	0.259**
			(0.090)	(0.115)
Constant	3.621***	3.530***	3.707***	3.735***
	(0.114)	(0.130)	(0.120)	(0.162)
Fixed effects				
Month	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes
Observations	71927	62823	71927	62823
Individuals	248	216	248	216
R^2	0.359	0.369	0.361	0.373

The table presents OLS estimates. Log net-revenues per sale are conditional on nonzero sales. Robust standard errors clustered on individual levels are in parentheses. In specifications 2 and 4, observations from individuals with an avg. RIR > 2 are excluded (n=28). In addition, for applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2. This exclusion ensures that there are at least two individuals per store. Observations are on an individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In specifications 3 and 4 of Table 2.4, we take the interactions between reciprocity and consulting intensity into account. In specification 3, when we include observations from all 248 salespersons, we see a positive but not significant interaction effect for both the medium- and the high-consulting-intensive area.[22] Interestingly and more importantly, in specification 4, in which we exclude hyper-reciprocal individuals with an average RIR higher than two, the interaction effect between reciprocity and the high-consulting-intensive area doubles in size and becomes significant on a 5% level. In contrast, in the low- and medium-consulting-intensive areas, reciprocity has no significant effect on net-revenues per sale.

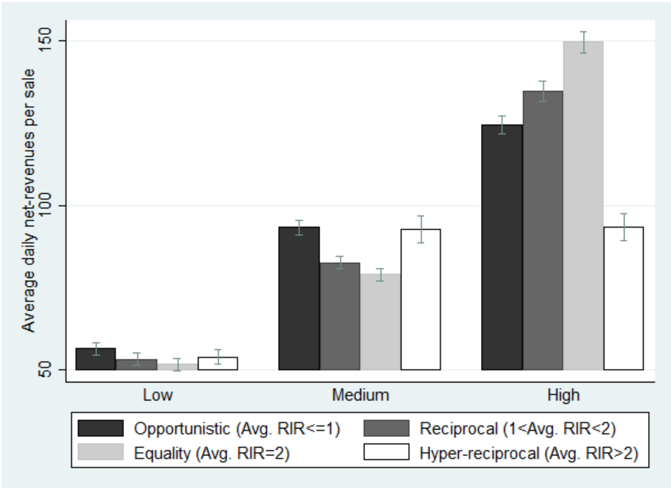
Besides the average RIR, we also observe that full-time employed salespeople achieve significantly higher net-revenues per sale than part-time employed salespeople. One explanation could be that part-time sales representatives have limited sales experience and less knowledge regarding the products. Furthermore, we find that male sales representatives generate more net-revenues per sale than female ones. Interestingly, when we additionally control for body height, results show that body height is the determining driver of higher daily net-revenues per sale rather than being male. Thus, taller salespeople performed significantly better. This is in line with the findings by Judge and Cable (2004), which show that body height is significantly related to workplace success. Regression models that include the individual's body height as a control variable are presented in Appendix B, in specifications 3 and 9 of Table 2.12. All our main effects remain significant in these specifications.

The economically and statistically significant finding that a salesperson's inclination towards reciprocal fairness is positively related to net-revenues per sale in the high-consulting-intensive area is also displayed in Figure 2.2. Here, we plot the mean net-revenues per sale for the behavioral types and consulting-intensive areas separately. It becomes apparent that independently of the behavioral type, daily net-revenues per sale increase with the consulting intensity. In addition, we can see that in the high-consulting-intensive area, the mean daily average net-revenues per sale increases with the average RIR, reaching the peak when participants equalized payoffs, and abruptly decreases for salespeople who behaved hyper-reciprocally in the trust game. This is in line with the results we find in specifications 3 and 4 of Table 2.4, showing that in the high-consulting-intensive area reciprocity is positively correlated with net-revenues per sale as long as hyper-reciprocal salespersons are excluded. In the medium- and low-consulting-intensive areas, differences in the mean daily average net-revenues per sale are small. To conclude, reciprocal salespeople generate higher revenues per sale in the high-consulting-intensive area where clients require thorough personal advice. Here, they can build on their strengths and convey the impression that they give valuable advice to

[22] When additionally controlling for salespersons' behavior in the role of the sender, demographics such as body height, education, and risk and time preferences, this effect becomes statistically significant (results available from the authors on request).

the customer and cross-sell related products, which in turn leads to higher revenues per customer.

FIGURE 2.2: Net-revenues per sale by behavioral types and consulting-intensive areas



The figure shows means and error bars.

2.4.5 Reciprocity, the number of sales and net-revenues

Given the positive relation between reciprocity and revenues per sale, the next question is whether an individual's willingness to reciprocate is profitable for the company. We first analyze the effect of an individual's willingness to reciprocate on the number of sales and then investigate the overall effect on total net-revenues. With regard to the number of sales, predictions can point in different directions. On the one hand, besides selling more per customer, reciprocal salespeople may also serve more customers successfully, due to their ability to convey trustworthiness. This would enhance the positive effect of reciprocity on total net-revenues. On the other hand, reciprocal salespeople may follow a more time-intensive sales strategy to truly understand the customer's needs. This would lead to a lower number of sales, and depending on whether the positive effect on net-revenues per sale or the negative effect on the number of sales has the greater impact, reciprocity can be either profitable or not.

In specifications 1 to 4 of Table 2.5, we estimate the association between the average RIR and the number of sales under the condition that the salesperson made at least one sale on the corresponding day. Specification 1 shows that reciprocity has a significant negative effect on the number of sales. In specification 2, when we exclude hyper-reciprocal individuals with an average RIR higher than two, the effect remains significant.[23] Additionally, the relation between the number of sales and our measurement of reciprocity is also quantitatively important. Specifications 1 and 2 show that compared to salespersons who returned nothing (average RIR = 0), an individual who returned the amount he or she received (average RIR = 1) serves approximately 13% and 15% fewer customers successfully, respectively. In specifications 3 and 4, we take the interaction effects between the consulting-intensive areas and reciprocity into account. Independently of whether individuals with an average RIR higher than two are excluded or not, reciprocity has a statistically significant and economically relevant negative effect on the number of sales in the area where customers require detailed advice. One potential intuition is as follows: in the high-consulting-intensive area, clients have to make more far-reaching decisions because items are more expensive and often not self-explanatory. Therefore, clients require thorough advice, which is, however, time-consuming. Reciprocal sales representatives may spend more time with each customer thoughtfully dealing with their concerns and offer the product alternative which best fits to their needs. However, spending more time with each customer also implies less time to get in contact with other clients and thus fewer sales.

We now consider the overall effect of reciprocity on net-revenues. This effect depends on whether the positive effect on net-revenues per sale or the negative effect on the number of sales is more pronounced. Specifications 5 to 8 of Table 2.5 show the OLS estimates for the relationship between reciprocity and net-revenues conditional on nonzero sales. Irrespectively of whether we exclude sales representatives with an average RIR higher than two or not, without the interaction effects, reciprocity has a negative but not statistically significant effect on daily net-revenues (specifications 5 and 6 of Table 2.5). Taking into consideration the interactions between consulting intensity and reciprocity, the relation between an individual's willingness to reciprocate and net-revenues is negative in the medium- and high-consulting-intensive areas. This effect is statistically significant on a 10% level in the high-consulting-intensive area independently of whether hyper-reciprocal individuals are excluded or not (specifications 7 and 8 of Table 2.5). However, when we measure reciprocity by the slope coefficient obtained by an OLS regression of the returns on the investments, the negative interaction effect between reciprocity and net-revenues becomes insignificant (see Appendix B, specifications 7 and 8 of Table 2.10). Thus, although reciprocal salespeople are able to sell more per customer, they are not able to generate higher total net-revenues. On the contrary, in some specifications, in the high-consulting-intensive area the negative effect of reciprocity on the number of

[23] As in specifications 2 and 4 of Table 2.4 where we exclude hyper-reciprocal salespersons, we additionally have to drop observations from four individuals with an average RIR smaller than or equal to two to ensure that we have at least two participants per store and we can apply store fixed effects.

sales even outweighs the positive effect on net-revenues per sale, resulting in marginally significantly lower net-revenues.

Additionally to the effects based on reciprocity, results show that tenure is a determining factor for daily net-revenues. As with all tenure estimates, this effect could be due to learning on the job or to sorting whether low-performing salespeople are dismissed or quit voluntarily. The tenure effect is additive and driven by both the number of sales and the net-revenues per sale. In addition, as men generate significantly more net-revenues per sale than women while serving a similar number of customers, they are able to achieve significantly higher total net-revenues. However, as already mentioned in Section 2.4.4, when additionally controlling for body height, the significant positive effect of being male disappears, indicating that body height matters instead of gender (see Appendix B, specifications 3 and 9 of Table 2.14).

As mentioned in Section 2.4.3, several salespeople were not able to sell anything on a given day. Therefore, a significant fraction of performance data has zero value. In Table 2.6, we present a probit model to test whether reciprocal salespeople are more likely to make a sale. The dependent variable is one when the salesperson made at least one sale on the corresponding day and zero otherwise. Regardless of considering the interactions between consulting intensity and reciprocity, all specifications of Table 2.6 suggest that reciprocity has no predictive power on the probability of making a sale. Furthermore, when excluding hyper-reciprocal individuals, the findings of Table 2.6 suggest that salespeople who returned something, even if they received nothing in the trust game, are more likely to make a sale. In addition, in specifications 1 and 2, salespeople who worked in the high-consulting-intensive area are more likely to make a sale. However, this effect disappears when we take the interaction effects into account. To summarizes, our findings confirm that the probability to enter into a successful sales interaction and sales performance in a given sales interaction are driven by different mechanisms. Whereas the probability of making a sale is independent of an individual's inclination toward reciprocal fairness, reciprocal behavior is an important predictor of sales performance conditional on a given sales interaction.

Table 2.5: Log net-revenues and log number of sales - OLS

	1	2	3	4	5	6	7	8
	Number of sales	Number of sales	Number of sales	Number of sales	Net-revenues	Net-revenues	Net-revenues	Net-revenues
	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions			Interactions	Interactions
Avg. RIR	-0.128**	-0.145**	0.019	0.074	-0.067	-0.053	0.029	0.049
	(0.060)	(0.072)	(0.099)	(0.098)	(0.051)	(0.060)	(0.090)	(0.093)
Return if investment = 0	-0.004	-0.001	-0.008	-0.004	-0.002	0.002	-0.005	0.000
	(0.010)	(0.011)	(0.010)	(0.010)	(0.008)	(0.009)	(0.008)	(0.009)
Medium consulting	0.170*	0.177*	0.380**	0.317	0.560***	0.581***	0.699***	0.622***
	(0.099)	(0.093)	(0.178)	(0.221)	(0.080)	(0.073)	(0.153)	(0.185)
High consulting	0.150	0.092	0.714***	0.796***	0.907***	0.896***	1.274***	1.243***
	(0.123)	(0.135)	(0.265)	(0.272)	(0.101)	(0.108)	(0.216)	(0.214)
Full-time	-0.105	-0.095	-0.105	-0.076	0.025	0.017	0.025	0.028
	(0.092)	(0.101)	(0.092)	(0.100)	(0.079)	(0.086)	(0.079)	(0.087)
Tenure in months	0.001**	0.000	0.001**	0.001	0.002***	0.001**	0.002***	0.001**
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.006	0.048	-0.023	0.012	0.198**	0.227**	0.180**	0.210**
	(0.110)	(0.111)	(0.111)	(0.109)	(0.088)	(0.090)	(0.091)	(0.090)
Medium consulting x avg. RIR			-0.136	-0.120			-0.090	-0.039
			(0.107)	(0.142)			(0.096)	(0.123)
High consulting x avg. RIR			-0.373**	-0.511***			-0.243*	-0.252*
			(0.166)	(0.193)			(0.131)	(0.145)
Constant	1.875***	2.126***	1.628***	1.733***	5.496***	5.656***	5.335***	5.468***
	(0.232)	(0.223)	(0.220)	(0.223)	(0.205)	(0.201)	(0.201)	(0.207)

Table 2.5: Continued

	1	2	3	4	5	6	7	8
	Number of sales	Number of sales	Number of sales	Number of sales	Net-revenues	Net-revenues	Net-revenues	Net-revenues
	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions			Interactions	Interactions
Fixed effects								
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71927	62823	71927	62823	71927	62823	71927	62823
Individuals	248	216	248	216	248	216	248	216
R^2	0.195	0.210	0.201	0.218	0.312	0.331	0.314	0.333

The table presents OLS estimates. Log number of sales and log net-revenues are conditional on nonzero sales. Robust standard errors clustered on individual levels are in parentheses. In specifications 1 to 4, we estimate the effect of the average RIR on the log number of sales, while in specifications 5 to 8, we estimate the effect of the average RIR on log net revenues. In specifications 2, 4, 6 and 8, observations from individuals with an avg. RIR > 2 are excluded (n=28). Furthermore, for applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2. This exclusion ensures that there are at least two individuals per store. Observations are on an individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.6: Probability of making a sale - probit model

	1	2	3	4
	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions
Avg. RIR	-0.128	0.046	0.045	0.022
	(0.133)	(0.171)	(0.165)	(0.272)
Return if investment = 0	0.011	0.043*	0.013	0.043*
	(0.022)	(0.024)	(0.022)	(0.024)
Medium consulting	-0.177	0.189	0.237	0.172
	(0.196)	(0.202)	(0.333)	(0.485)
High consulting	0.604*	0.834**	0.794	0.696
	(0.318)	(0.362)	(0.811)	(0.840)
Full-time	-0.069	-0.294	-0.051	-0.296
	(0.185)	(0.201)	(0.185)	(0.206)
Tenure in months	0.002	0.001	0.002	0.001
	(0.001)	(0.002)	(0.001)	(0.001)
Male	-0.048	0.062	-0.038	0.075
	(0.223)	(0.222)	(0.230)	(0.239)
Medium consulting x avg. RIR			-0.269	0.012
			(0.190)	(0.363)
High consulting x avg. RIR			-0.127	0.100
			(0.415)	(0.515)
Constant	1.210*	0.394	0.998	0.435
	(0.628)	(0.634)	(0.664)	(0.705)
Fixed effects				
Month	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes
Observations	80236	69395	80236	69395
Individuals	247	215	247	215
Pseudo R^2	0.295	0.346	0.297	0.346
Wald Chi^2	1981.73	8231.07	2139.74	9028.31

The table presents probit estimates. Robust standard errors clustered on individual levels are in parentheses. In specifications 2 and 4, observations from individuals with an avg. RIR > 2 are excluded (n=28). Furthermore, for applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2. This exclusion ensures that there are at least two individuals per store. Observations of one individual are dropped because of perfectly predicted success. Observations are on an individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.4.6 Robustness analysis

Below we probe the robustness of our results using different approaches. First, we test whether our results are robust by using different measurements of reciprocity. In Appendix B, in Tables 2.9, 2.10, and 2.11, we replace the average RIR with a reciprocity measurement obtained by running simple OLS regressions of returns on the investments. The slope coefficient represents an individual's inclination towards reciprocity (RC) and further accounts for the case when the investment is zero. In specifications 1 and 2 of Table 2.9, we find a positive correlation between the RC and net-revenues per sale. This result is consistent with our findings when we used the average RIR. When we exclude hyper-reciprocal individuals and measure reciprocity by the RC instead of the average RIR, the result that reciprocal individuals sell significantly more per customer in the area where advice really matters is, however, only significant on a 10% level rather than on a 5% level. In specifications 1 and 2 of Table 2.10, the coefficients of the RC variable and in specifications 3 and 4 of Table 2.10, the interaction effects between the high-consulting-intensive area and the RC show a negative impact on the number of sales. The coefficients remain significant and have the same sign as those in Table 2.5 where reciprocity is measured by the average RIR. While the previously mentioned results are qualitatively robust, regressions estimating net-revenues show that the interaction effect between reciprocity and high-consulting-intensity becomes insignificant when using the RC instead of the average RIR (see Appendix B, specifications 7 and 8 of Table 2.10). Lastly, we can show that the coefficients of the probit regressions reported in Table 2.6 do not change qualitatively when we apply the RC instead of the average RIR (see Appendix B, Table 2.11). We run several other robustness checks:[24] we estimate a salesperson's willingness to reciprocate by regressing the back transfers on the investments forcing the slope through the origin. Results are consistent with the regressions using the RC that also considers the return when the investment is zero. In addition, we rerun all specifications using a dummy variable that is one for individuals with an average RIR higher than one and zero otherwise. Results confirm that salespersons with an average RIR higher than one sell significantly more per customer, have a lower number of sales, and lower net-revenues. Additionally, we re-estimate all regressions without controlling for salespeople's back transfer when the investment was zero. Again, estimates are statistically and economically robust.

Next, to assess whether our findings are robust and not driven by other sales representatives' characteristics, we additionally include control variables measuring trusting behavior, an individual's education, body height, risk and time preferences, and the Big 5 personality dimensions. We present the robustness analysis with the inclusion of

[24] Results available from the authors on request.

other control variables for two of our main specifications. First, we probe the robustness for the specification that includes observations from all salespersons independently of their average RIR (specification 1), and second, for the specification in which individuals with an average RIR higher than two are excluded and interaction effects between the consulting-intensive area and reciprocity are considered (specification 4). Results are qualitatively the same if we consider specifications 2 and 3. In Appendix B, Table 2.12 presents the results for net-revenues per sale, Table 2.13 for the number of sales, Table 2.14 for net-revenues, and Table 2.15 for the probability of making a sale.

In specification 2, we control for an individual's trusting behavior, measured by the investment in the trust game. Because we limited the sender's choice set to four amounts (0, 6, 12, and 18 Euro), trusting behavior is expressed by an indicator variable with four categories. Salespeople who sent nothing represent the omitted category. Our findings show that the inclusion of trusting behavior has no considerable effect on the magnitude and statistical significance of the average RIR coefficient in either the regression models in which net-revenues per sale or the number of sales are estimated. This indicates that reciprocity is determining for sales performance even when we control for trusting behavior. However, when controlling for trusting behavior, we find a negative and marginally significant effect of the average RIR on net-revenues even when we do not consider the interaction effect between reciprocity and consulting intensity (see Appendix B, specification 2 of Table 2.14). In addition, results show that the behavior in the role of the sender has neither a predictive power for net-revenues per sale nor for the number of sales. While the trusting behavior has no significant effect on net-revenues when observations from all salespersons are included (see Appendix B, specification 2 of Table 2.14), we find that salespeople who transferred 6 Euro achieved significantly less net-revenues than those who sent nothing when we exclude hyper-reciprocal individuals and consider the interaction effects between reciprocity and consulting intensity (see Appendix B, specification 8 of Table 2.14). Furthermore, specification 2 of Table 2.15 reveals that controlling for trusting behavior leads to a significant negative effect of reciprocity on the probability of making a sale. In addition, specification 2 of Table 2.15 suggests that salespeople who transferred either 6 Euro or their total initial endowment in the role of the sender are significantly more likely to make a sale. If trusting behavior measures an individual's confidence in relying on another person, one explanation for this result could be that individuals who transferred money in the role of the sender may be those who take the first step in approaching a potentially lucrative sales interaction. However, in specification 8 of Table 2.15, when we exclude hyper-reciprocal individuals and consider the interactions between consulting intensity and reciprocity, the negative effect of reciprocity as well as the positive effects of sending 6 or 18 Euro become insignificant.

Given the importance of individual heterogeneity for sales performance, we also test how sensible our findings are when adding two further important demographics: body height and education. Body height is measured in centimeters, and education is represented by an indicator variable with the following three categories: lower education (secondary school diploma and/or apprenticeships), higher education (high school diploma and/or university degree), and other education. The category lower education serves as reference group. Both body height and education are assumed to be not only determining factors for sales success but may also correlate with reciprocity and thus can be responsible for the estimated effects. For example, Dohmen et al. (2008) provided evidence that taller people are more reciprocal than smaller ones. In addition, previous research has shown that body height has a positive effect on workplace success and income (Judge and Cable 2004). Evidence on whether or not education is associated with reciprocity is mixed. While Bellemare and Kröger (2007) found that individuals with a lower level of education behaved more reciprocally in the trust game than more highly educated persons, Fehr et al. (2003) did not find an education effect. Further, as education is an indicator for skills, it may also drive sales performance. Tables 2.12, 2.13, 2.14, and 2.15 in Appendix B show that including these demographic characteristics has no considerable effect on the magnitude and statistical significance of the reciprocity coefficients. While the nature of education has no predictive power for sales performance, body size has a significant positive effect on daily average net-revenues per sale and in turn also for daily total net-revenues (see Appendix B, specifications 3 and 9 of Table 2.12 and Table 2.14). As mentioned in Sections 2.4.4 and 2.4.5, when controlling for body height, the significant positive effect of being male disappears, indicating that body height matters rather than gender.

In addition, we assess how robust our results are when we control for risk and time preferences. Both of these factors might influence sales strategies. For example, a risk averse salesperson may be tempted to uphold current selling techniques, while a risk seeking individual may try out new strategies. Further, patient salespeople may exert less pressure to close a sale than impatient ones. We observe that risk preferences are significantly associated with net-revenues (see Appendix B, specifications 4 and 10 of Table 2.14). The more risk-seeking a salesperson is, the lower the net-revenues he or she generates. We do not find a significant association between time preferences and sales performance variables. In addition, we do not find that either risk or time preferences have an effect on the magnitude or the significance level of the average RIR coefficient.

As a final robustness check, we test if our results are robust to the inclusion of the Big 5 personality dimensions. We elicited a salesperson's personality traits by a short version of the Big 5 inventory (Gerlitz and Schupp 2005). This test measures five major dimensions of personality: conscientiousness, extraversion, agreeableness, openness to new experiences, and neuroticism. Previous research shows that personality traits are not only correlated with job performance (Barrick and Mount 1991, Vinchur et al. 1998) but also have predictive power for reciprocity (Dohmen et al. 2008). Dohmen et al. (2008) showed that all five personality dimensions are positively and significantly related to reciprocity. In particular, they found that conscientiousness and agreeableness have the biggest impact. As our results show, the reciprocity coefficients and interactions between reciprocity and consulting intensity stay significant when we include the five personality types (see Appendix B, specifications 5 and 11 of Tables 2.12 and 2.13, and specification 11 of Table 2.14). While conscientiousness has no significant effect on the net-revenues per sale and the number of sales, we find a positive and marginally significant association between conscientiousness and the net-revenues (see Appendix B, specification 5 of Table 2.14). This result is consistent with the findings by Vinchur et al. (1998) which revealed that conscientiousness is positively correlated to sales performance. In addition, our results show that conscientious salespeople are less likely to make a sale when hyper-reciprocal individuals are excluded (see Appendix B, specification 11 of Table 2.15). Furthermore, we find a negative and significant effect of neuroticism on the net-revenues per sale (see Appendix B, specifications 5 and 11 of Table 2.12). This seems reasonable because neurotic salespeople might be less convincing and less able to smoothly interact with their customers. In addition, we also find a negative and marginally significant relationship between agreeableness and net-revenues (see Appendix B, specifications 5 and 11 of Table 2.14). As agreeable salespersons seem to be great team players, they might sacrifice their sales target for the success of their co-workers in a competitive sales environment, like ours. This is in line with Judge et al. (1999), who argued that agreeableness is negatively associated with extrinsic career success.

Finally, we present the full model including all control variables in specifications 6 and 12 of Tables 2.12, 2.13, 2.14, and 2.15 in Appendix B. Even when adding all control variables at once, our estimates are qualitatively robust. To sum up, we can show that our results are robust to alternative reciprocity measurements and to the inclusion of different controls.

2.5 Conclusion

Recent research suggests that pro-social characteristics are determining for individual work performance and earnings (Bowles et al. 2001, Barr and Serneels 2009, Leibbrandt 2012). In this paper we examine the direct relationship between a salesperson's willingness to reciprocate and sales performance in real-world sales interactions. By combining an experimental measurement of reciprocity with a unique field data set on individual sales performance, we find in a first step that reciprocal salespeople sell more per customer. In particular, we show that especially in the product area of high-consulting-intensity, salespeople with an inclination towards reciprocal fairness achieved higher net-revenues per sale. Reciprocal and equality-minded sales representatives may be more attentive to the customers' needs and may try to find the best solution for their clients without losing sight of their own and the company's interests. We further believe that they are more able to signal their trustworthiness and create a positive feeling about the store. As a result, customers trust them and buy more. In particular, we assume that they are more able to cross-sell products that are of interest to the customer. In addition, our data shows that hyper-reciprocal individuals, who sent back more than the amount that would equalize payoffs, generate, on average, the lowest net-revenues per sale. In line with Grant (2013), we suggest that these poor-performing salespeople may be more focused on their customers' needs than on their sales targets. In a second step and most importantly from a company's point of view, we investigate whether reciprocity is profitable for the enterprise. We find that reciprocal salespersons make fewer sales closures. In the area where clients require thorough advice, this negative effect of reciprocity on the number of sales even outweighs the positive effect of an individual's willingness to reciprocate on net-revenues per sale, resulting in marginally significantly lower net-revenues. One explanation for this result may be that reciprocal sales representatives engage in more extensive and time-consuming sales talks, implying less time for other customers.

Our findings not only enable us to obtain deeper insights into the relationship between reciprocity and sales success but also show that the reciprocal behavior observed in the trust game is an important predictor of real-world sales settings, although there are contextual differences between the experimental game and the field environment. In our trust game, salespersons were anonymously paired with persons from Austria who do not work for the company, while in the field they interact face-to-face with their customers. Previous research shows that people often trust strangers based on factors such as perceived attractiveness, gender, facial similarity, and race (e.g., DeBruine 2002, Chaudhuri and Gangadharan 2003, Eckel and Wilson 2003, Wilson and Eckel 2006). Whereas these factors were not part of the experimental game, they are present in real-life sales interactions and may shape a customer's willingness to trust. An improved understanding of the role of stereotypes in the sales context and whether stereotypically trustworthy salespeople are more successful seems to be an interesting topic for future research. Moreover, we implemented a one-shot trust game, whereas repeated customer interactions and customer-to-customer recommendations are common in natural sales environments. Through long-term customer-relationships and a spotless reputation, trust can be built up and sustained. Furthermore, it might well be that a customer who has dealt with a reciprocal salesperson is more satisfied with the selling process and more likely to return. List (2006) even suggests that in the marketplace pro-social behavior mainly occurs because of reputational concerns. Leibbrandt (2012), on the other hand, shows that pro-sociality, measured by a one-shot public goods game, is associated with stable and longer-lasting trade relations. More research is certainly warranted to advance our understanding of the long-term effect of reciprocity on sales performance and customer satisfaction. In addition, related work suggests that pro-sociality in the workplace is able to establish an environment where customers get the impression that salespeople care about their needs (Podsakoff et al. 2009). Therefore, it would also be worth examining whether business units with a higher share of reciprocal employees are more successful and whether pro-social salespeople behave reciprocally not just towards clients but also towards their co-workers even in a competitive sales setting, like ours.

We believe that our results have high practical relevance. We show that experimental measurements provide a simple way to identify characteristics of successful salespersons. Furthermore, we provide evidence that a reciprocal individual applies a sales strategy other than an opportunistic salesperson. While data suggests that reciprocal individuals are good at building trust and convincing the customer to buy more, opportunistic individuals are more efficient at closing a sale and proceeding with the next customer. Thus, based on our results, we can propose the following management implications. First, knowing the characteristics of their sales force helps managers to consider individual strengths by assigning tasks. Second, understanding what characteristics it takes to achieve sales targets in different sales environments helps to simplify and accelerate the recruiting process as well as to select proper and individualized training. For example, managers can use experimental games to assess important characteristics and place the person with the appropriate traits in the position to be filled. In addition, as reciprocal salesperson may serve fewer customers successfully due to time constraints, they can undergo training in time management. Third, knowing the qualities of the sales staff may also enable managers to design incentive schemes that counteract the shortcomings of both the reciprocal and the opportunistic type.

2.6 References

Akerlof, G. A. (1982). Labor contracts as a partial gift exchange. *Quarterly Journal of Economics*, 97(4):543–569.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2008). Using selection on observed variables to assess bias from unobservables when evaluating Swan-Ganz Catheterization. *American Economic Review*, 98(2):345–350.

Baran, N. M., Sapienza, P., and Zingales, L. (2010). Can we infer social preferences from the lab? Evidence from the trust game. *NBER Working Paper 15654*.

Barr, A. and Serneels, P. (2009). Reciprocity in the workplace. *Experimental Economics*, 12(1):99–112.

Barrick, M. R. and Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1):1–26.

Bellemare, C. and Kröger, S. (2007). On representative social capital. *European Economic Review*, 51(1):183–202.

Bellows, J. and Miguel, E. (2009). War and local collective action in Sierra Leone. *Journal of Public Economics*, 93(11-12):1144–1157.

Ben-Ner, A. and Putterman, L. (2001). Trusting and trustworthiness. *Bosten University Law Review*, 81:523–551.

Benz, M. and Meier, S. (2008). Do people behave in experiments as in the field? - Evidence from donations. *Experimental Economics*, 11(3):268–281.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Boly, A. (2011). On the incentive effects of monitoring: Evidence from the lab and the field. *Experimental Economics*, 14(2):241–253.

Bowles, S., Gintis, H., and Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature*, 39(4):1137–1176.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398.

Burks, S. V., Carpenter, J. P., and Verhoogen, E. (2003). Playing both roles in the trust game. *Journal of Economic Behavior & Organization*, 51(2):195–216.

Caldwell, C. and Clapham, S. (2003). Organizational trustworthiness: An international perspective. *Journal of Business Ethics*, 47(4):349–364.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. The roundtable series in behavioral economics. Russell Sage Foundation and Princeton University Press, New York, N.Y. and Princeton, N.J.

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics using Stata*. Stata Press, College Station, Texas, rev. ed. 2010 edition.

Carlsson, F., Johansson-Stenman, O., and Nam, P. K. (2014). Social preferences are stable over long periods of time. *Journal of Public Economics*, 117:104–114.

Carpenter, J. and Seki, E. (2011). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, 49(2):612–630.

Carter, M. R. and Castillo, M. (2002). The economic impacts of altruism, trust and reciprocity: An experimental approach to social capital. *Wisconsin-Madison Agricultural and Applied Economics Staff Papers 448*.

Chaudhuri, A. and Gangadharan, L. (2003). Gender differences in trust and reciprocity. *Working Paper unpublished*.

Cleave, B., Nikiforakis, N., and Slonim, R. (2013). Is there selection bias in laboratory experiments? The case of social and risk preferences. *Experimental Economics*, 16(3):372–382.

Cooper, D. J. and Kagel, J. H. (2013). Other-regarding preferences: A selective survey of experimental results. In Kagel, J. H. and Roth, A. E., editors, *forthcoming in The Handbook of Experimental Economics*, volume 2. Princeton university press, Princeton.

Cox, J. C. (2002). Trust, reciprocity, and other-regarding preferences: Groups vs. individuals and males vs. females. In *Experimental Business Research*, pages 331–350. Springer.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2):260–281.

Dearmon, J. and Grier, K. (2009). Trust and development. *Journal of Economic Behavior & Organization*, 71(2):210–220.

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1498):1307–1312.

Dincer, O. C. and Uslaner, E. M. (2010). Trust and growth. *Public Choice*, 142(1-2):59–67.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2008). Representative trust and reciprocity: Prevalence and determinants. *Economic Inquiry*, 46(1):84–90.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2009). Homo reciprocans: Survey evidence on behavioural outcomes. *Economic Journal*, 119(536):592–612.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Dow, W. H. and Norton, E. C. (2003). Choosing between and interpreting the Heckit and two-part models for corner solutions. *Health Services and Outcomes Research Methodology*, 4(1):5–18.

Eckel, C. C. and Wilson, R. K. (2003). The human face of game theory: Trust and reciprocity in sequential games. In Ostrom, E. and Walker, J., editors, *Trust and reciprocity: Interdisciplinary lessons from experimental research*, pages 245–274. Russel Sage Foundation, New York.

Falk, A., Meier, S., and Zehnder, C. (2013). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association*, 11(4):839–852.

Fehr, E., Fischbacher, U., Rosenbladt, B. v., Schupp, J., and Wagner, G. G. (2003). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *IZA Discussion paper 715*.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics*, 95(9):1144–1155.

Fehr, E. and List, J. A. (2004). The hidden costs and returns of incentives: Trust and trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5):743–771.

Franke, G. R. and Park, J.-E. (2006). Salesperson adaptive selling behavior and customer orientation: A meta-analysis. *Journal of Marketing Research*, 43(4):693–702.

Fukuyama, F. (2001). Social capital, civil society and development. *Third World Quarterly*, 22(1):7–20.

Gerlitz, J.-Y. and Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persoenlichkeitsmerkmale im SOEP. *DIW Research Notes 4*.

Grant, A. M. (2013). *Give and take: Why helping others drives our success.* Penguin Group, New York.

Guiso, L., Sapienza, P., and Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, 94(3):526–556.

Güth, W., Huck, S., and Müller, W. (2001). The relevance of equal splits in ultimatum games. *Games and Economic Behavior*, 37(1):161–169.

Hardin, R. (2002). *Trust and trustworthiness.* Russell Sage Foundation, New York.

Homburg, C., Müller, M., and Klarmann, M. (2011). When should the customer really be king? On the optimum level of salesperson customer orientation in sales encounters. *Journal of Marketing*, 75(2):55–74.

Jobber, D. and Lancaster, G. (2006). *Selling and sales management.* Financial Times/Prentice Hall, Harlow, England and New York, 7th ed. edition.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.

Judge, T. A. and Cable, D. M. (2004). The effect of physical height on workplace success and income: Preliminary test of a theoretical model. *Journal of Applied Psychology*, 89(3):428–441.

Judge, T. A., Higgins, C. A., Thoresen, C. J., and Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, 52(3):621–652.

Karlan, D. S. (2005). Using experimental economics to measure social capital and predict financial decisions. *American Economic Review*, 95(5):1688–1699.

Knack, S. and Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics*, 112(4):1251–1288.

Kosfeld, M. and Rustagi, D. (2015). Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia. *American Economic Review*, 105(2):747–783.

Lammers, F. (2010). Fairness in delegated bargaining. *Journal of economics & management strategy*, 19(1):169–183.

Leibbrandt, A. (2012). Are social preferences related to market performance? *Experimental Economics*, 15(4):589–603.

Levitt, S. D. and List, J. A. (2007a). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics*, 40(2):347–370.

Levitt, S. D. and List, J. A. (2007b). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174.

List, J. A. (2006). The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political Economy*, 114(1):1–37.

Madden, D. (2008). Sample selection versus two-part models revisited: The case of female smoking and drinking. *Journal of Health Economics*, 27(2):300–307.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

Oxoby, R. J. and McLeish, K. N. (2004). Sequential decision and strategy vector methods in ultimatum bargaining: Evidence on the strength of other-regarding behavior. *Economics Letters*, 84(3):399–405.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., and Blume, B. D. (2009). Individual- and organizational-level consequences of organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology*, 94(1):122–141.

Putnam, R. D., Leonardi, R., and Nanetti, R. Y. (1994). *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.

Rustagi, D., Engel, S., and Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, 330(6006):961–965.

Saxe, R. and Weitz, B. A. (1982). The SOCO scale: A measure of the customer orientation of salespeople. *Journal of Marketing Research*, 19(3):343–351.

Schechter, L. (2007). Traditional trust measurement and the risk confound: An experiment in rural Paraguay. *Journal of Economic Behavior & Organization*, 62(2):272–292.

Serra, D., Serneels, P., and Barr, A. (2011). Intrinsic motivations and the non-profit health sector: Evidence from Ethiopia. *Personality and Individual Differences*, 51(3):309–314.

Stoop, J., Noussair,, Charles N., and van Soest, D. (2012). From the lab to the field: Cooperation among fishermen. *Journal of Political Economy*, 120(6):1027–1056.

Vinchur, A. J., Schippmann, J. S., Switzer, Fred S., III, and Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology*, 83(4):586–597.

Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., and Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. *Economics Letters*, 120(2):142–145.

Wilson, R. K. and Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, 59(2):189–202.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge, Mass., 2nd ed. edition.

2.7 Appendix A: Experimental instructions - general rules and instructions for the trust game

(Original instructions are in German)

General rules of the game

Thank you very much for participating in this research project. In the following we will explain the general rules of the game:

- ✓ The participation documents include five games and a short questionnaire.
- ✓ Each fourth participant will be paid out according to his/her decisions and the decision of other players.
- ✓ The draw decides which participants and which games will be paid out.
- ✓ Only completely filled-in documents will be considered for the payment.
- ✓ All information will be treated in strict confidence and anonymity.
- ✓ As we are interested in the personal choices of each participant, it is very important to us **that every participant makes his or her own decisions alone at home. An agreement with your family, your friends or colleagues not only spoils the game for yourself, but also harms our research project.**

Thank you very much for your cooperation!

Participation is as simple as that:

1. Please read the instructions and the example for the first game.
2. Please answer then the corresponding comprehension questions and indicate your decisions on the red-rimmed pages.
3. Afterwards you can continue with the next game.
4. Once you have finished all the games, please fill in the short questionnaire.
5. Please return the fully completed documents using the enclosed envelope. Closing date is **November 30, 2013.**

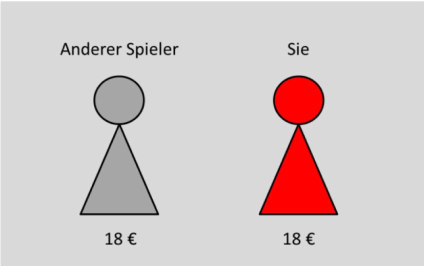
If you have any questions concerning the rules of the game, please contact our colleague xxx by phone on xxx (we are happy to call you back) or by e-mail at xxx.

Thank you very much for your participation. We wish you a lot of fun with the different games!

Game 2: What is being done…

We will now describe game 2. Please read through the instructions carefully.

1) Start



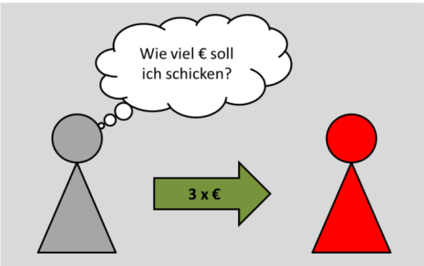
There are 2 players: You and another person.

You are the red player.

The other person is a real person, he/she lives in Austria, but doesn't work for the company XX.

Each player receives an initial endowment of 18 €.

2) Decision of the other player

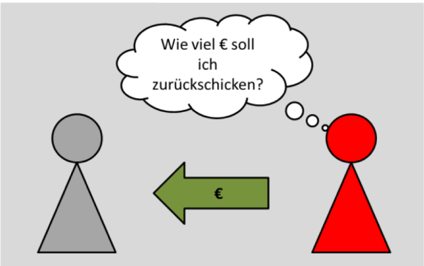


The other player decides first. He/she can send you either 0, 6, 12 or 18 € of his 18 €.

The money he/ she sends to you is taken away from him/her.

At the same time, each € the other player sends to you is ***tripled***. This means that you receive three times the amount the other player has sent to you.

3) Your decision



You see how much money the other player has sent to you and you can now send money back to him/her.

The money you send back to the other player is removed from your account.

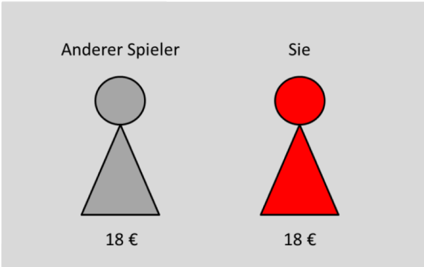
However, the amount ***will not*** be tripled this time. The other player receives exactly as many € as you send to him/her.

What do you receive? You receive the initial endowment of 18 €, ***plus*** triple the amount that the other player sends to you, ***minus*** the money you send back to the other player.

What does the other player receive? He/she receives whatever he/she keeps of his/her initial endowment of 18 €, ***plus*** the amount you send back to him/her.

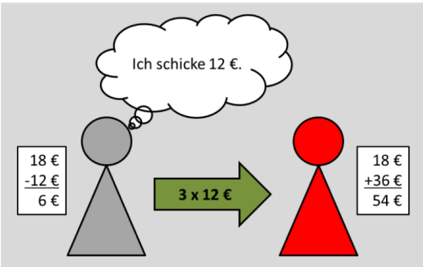
Example:

1) Start



You and the other player each receive 18 €.

2) Decision of the other player

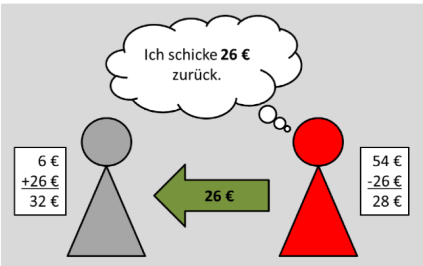


The other player sends you 12 €.

The money is taken away from him/her.

The 12 € is tripled. This means that you additionally receive 36 €. At this point the other player has 6 € and you have 54 €.

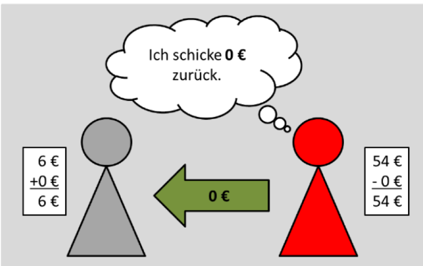
3) Your decision



You can send back to the other player between 0 and 54 €.

Let's assume that you decide to send back 26 €.

Then the other player receives 32 € at the end and you receive 28 €.



Let's assume that you decide to send back 0 €.

Then the other player receives 6 € at the end and you receive 54 €.

ID

Comprehension questions:

To ensure that we have explained game 2 comprehensibly, we kindly ask you to briefly answer the following comprehension questions:

Please assume that:

- You and the other player start with 18 €.
- The other player sends you 6 €.
- This 6 € is tripled. Thus, you receive additionally to your initial endowment of 18 € again 18 € (3 x 6 = 18 €). At this point you have 36 € and the other player has 12 €.
- You can now send back between 0 and 36 €. You decide to send back 6 €.

Please check the right answer:

How much money do you have in this case at the end of the game?

☐ 12 €

☐ 20 €

☐ 30 €

How much money does the other player have at the end of the game?

☐ 12 €

☐ 18 €

☐ 24 €

ID

Decision sheet 2

As a reminder: Please remember that your decision is about real money. Each fourth employee of the company XX who takes part in our study will be paid out in cash for one of the games. The other player will be paid out in cash as well, when this game has been drawn. The payment takes place anonymously, i.e. no one finds out what you or the other player has received.

You have an initial endowment of 18 €:

The other player makes his/her choice at the same time as you. As you do not yet know how much money the other player will send you, we ask you to make a decision for all 4 possible options of the other player:
(Please fill in the amount in Euro **in each of the 4** gray boxes.)

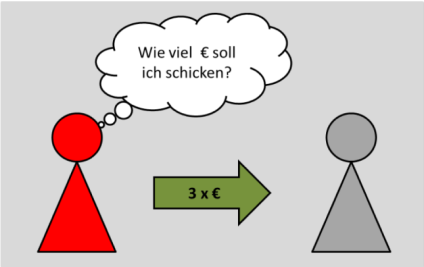
Assume the other player sends you…	**How much money do you then send back to the other player?**
0 €, i.e., you would have 18 €. The other player would have 18 €.	**…………€** [number between 0 and 18]
6 €, i.e., you would have 36 €. The other player would have 12 €.	**…………€** [number between 0 and 36]
12 €, i.e., you would have 54 €. The other player would have 6 €.	**…………€** [number between 0 and 54]
18 €, i.e., you would have 72 €. The other player would have 0 €.	**…………€** [number between 0 and 72]

Game 3: What is being done…

You play again with the same person from game 2. The rules of the game are the same as in game 2.

The only change is that the roles are changed. Now it is up to you to decide first how much money you want to send to the other player:

Your decision



You are now first and you can decide how much of your 18 €, either 0, 6, 12 or 18 €, you want to send to the other player.

What do you receive? You receive whatever you keep of your initial endowment of 18 €, *plus* the amount the other player sends back to you.

What does the other player receive? He/she receives his initial endowment of 18 €, *plus* the tripled amount you send him/her, *minus* the money that he/she sends back to you.

ID

Decision sheet 3

As a reminder: Please remember that your decision is about real money. Each fourth employee of the company XX who takes part in our study will be paid out in cash for one of the games. The other player will be paid out in cash as well, when this game has been drawn. The payment takes place anonymously, i.e. no one finds out what you or the other player has received.

You have an initial endowment of 18 €:

How much money do you send to the other player?
(Please check.)

☐ 0 €

☐ 6 €

☐ 12 €

☐ 18 €

2.8 Appendix B: Further results

Table 2.7: Selection effects (all stores): probit regression

	1 Benchmark	2 Net-revenues per sale	3 Net-revenues	4 Number of sales	5 Fraction of days with a sale
Male	-0.194**	-0.193**	-0.224**	-0.195**	-0.194**
	(0.096)	(0.096)	(0.098)	(0.096)	(0.096)
Flagship	-0.212***	-0.211***	-0.212***	-0.195**	-0.212***
	(0.077)	(0.077)	(0.077)	(0.078)	(0.077)
Medium consulting	-0.026	-0.026	-0.115	-0.047	-0.026
	(0.097)	(0.097)	(0.102)	(0.099)	(0.097)
High consulting	-0.062	-0.060	-0.232*	-0.074	-0.061
	(0.122)	(0.122)	(0.135)	(0.122)	(0.121)
Full-time	-0.065	-0.066	-0.075	-0.063	-0.064
	(0.094)	(0.094)	(0.095)	(0.094)	(0.095)
Tenure in months	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Net-revenues per sale		-0.000			
		(0.000)			
Net-revenues			0.000***		
			(0.000)		
Number of sales				0.007	
				(0.005)	
Fraction of days with a sale					-0.025
					(0.199)
Constant	-0.525***	-0.524***	-0.617***	-0.608***	-0.505**
	(0.112)	(0.112)	(0.116)	(0.127)	(0.200)
Observations	1369	1369	1369	1369	1369
Pseudo R^2	0.014	0.014	0.021	0.015	0.014
Wald Chi^2	19.36	19.43	28.55	21.59	19.36

The table reports probit estimates. The dependent variable equals one if the sales representative participated in the study and zero otherwise. Data includes observations from individuals who received the study documents and worked as a sales representative for at least one month between March 2012 and March 2014. Robust standard errors clustered on individuals are in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.8: Selection effects (stores with more than one participant): probit regression

	1 Benchmark	2 Net-revenues per sale	3 Net-revenues	4 Number of sales	5 Fraction of days with a sale
Male	-0.150	-0.147	-0.176*	-0.152	-0.149
	(0.104)	(0.104)	(0.106)	(0.104)	(0.104)
Flagship	-0.287***	-0.285***	-0.286***	-0.269***	-0.288***
	(0.084)	(0.084)	(0.084)	(0.085)	(0.084)
Medium consulting	-0.011	-0.009	-0.088	-0.036	-0.011
	(0.106)	(0.106)	(0.112)	(0.108)	(0.106)
High consulting	-0.027	-0.018	-0.170	-0.039	-0.024
	(0.131)	(0.132)	(0.144)	(0.131)	(0.131)
Full-time	-0.082	-0.083	-0.090	-0.077	-0.078
	(0.102)	(0.102)	(0.103)	(0.102)	(0.103)
Tenure in months	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Net-revenue per sale		-0.000			
		(0.000)			
Net-revenue			0.000**		
			(0.000)		
Number of sales				0.009	
				(0.006)	
Fraction of days with a sale					-0.077
					(0.227)
Constant	-0.508***	-0.502***	-0.590***	-0.612***	-0.443**
	(0.121)	(0.121)	(0.125)	(0.138)	(0.225)
Observations	1165	1165	1165	1165	1165
Pseudo R^2	0.016	0.016	0.021	0.018	0.016
Wald Chi^2	18.92	19.69	24.55	21.54	18.97

The table reports probit estimates. The dependent variable equals one if the sales representative participated in the study and zero otherwise. Data includes observations from individuals who received the study documents, worked in stores with more than one participating salesperson and worked as a salesperson for at least one month between March 2012 and March 2014. Robust standard errors clustered on individuals are in parentheses. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.9: Log net-revenues per sale - RC with intercept

	1	2	3	4
	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions
RC	0.060**	0.079**	0.007	-0.018
	(0.028)	(0.035)	(0.046)	(0.064)
Medium consulting	0.398***	0.414***	0.306***	0.314**
	(0.049)	(0.053)	(0.109)	(0.126)
High consulting	0.762***	0.815***	0.646***	0.599***
	(0.065)	(0.071)	(0.126)	(0.144)
Full-time	0.142***	0.124**	0.141***	0.115**
	(0.048)	(0.056)	(0.047)	(0.055)
Tenure in months	0.001	0.001	0.001	0.001
	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.175***	0.160**	0.177***	0.174**
	(0.060)	(0.068)	(0.059)	(0.067)
Medium consulting x RC			0.068	0.086
			(0.063)	(0.083)
High consulting x RC			0.088	0.177*
			(0.077)	(0.094)
Constant	3.663***	3.531***	3.735***	3.673***
	(0.120)	(0.128)	(0.130)	(0.155)
Fixed effects				
Month	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes
Observations	71927	62823	71927	62823
Individuals	248	216	248	216
R^2	0.359	0.369	0.360	0.371

The table presents OLS estimates. Log net-revenues per sale are conditional on nonzero sales. Reciprocity is measured by the slope coefficient of a regression of back transfers on investments. Robust standard errors clustered on individual levels are in parentheses. In specifications 2 and 4, observations from individuals with an average RIR > 2 are excluded (n=28). Furthermore, for applying store fixed effects, we have to exclude four additional individuals with an avg. RIR $<= 2$. This exclusion ensures that there are at least two individuals per store. Observations are on individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.10: Log net-revenues and log number of sales - RC with intercept

	1	2	3	4	5	6	7	8
	Number of sales	Number of sales	Number of sales	Number of sales	Net-revenues	Net-revenues	Net-revenues	Net-revenues
	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions			Interactions	Interactions
RC	-0.097*	-0.113*	0.056	0.071	-0.037	-0.034	0.063	0.053
	(0.057)	(0.058)	(0.101)	(0.092)	(0.048)	(0.049)	(0.089)	(0.078)
Medium consulting	0.151	0.162*	0.384**	0.327*	0.548***	0.575***	0.690***	0.641***
	(0.097)	(0.091)	(0.170)	(0.171)	(0.079)	(0.072)	(0.147)	(0.143)
High consulting	0.133	0.074	0.516**	0.506**	0.896***	0.890***	1.162***	1.105***
	(0.122)	(0.132)	(0.231)	(0.225)	(0.102)	(0.109)	(0.190)	(0.177)
Full-time	-0.122	-0.105	-0.119	-0.087	0.021	0.018	0.023	0.028
	(0.093)	(0.103)	(0.092)	(0.102)	(0.078)	(0.087)	(0.078)	(0.087)
Tenure in months	0.001*	0.000	0.001*	0.001	0.002***	0.001**	0.002***	0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.039	0.072	0.027	0.043	0.214**	0.232***	0.204**	0.217**
	(0.107)	(0.110)	(0.107)	(0.110)	(0.086)	(0.088)	(0.087)	(0.089)
Medium consulting x RC			-0.172	-0.145			-0.104	-0.059
			(0.112)	(0.115)			(0.102)	(0.099)
High consulting x RC			-0.289*	-0.354**			-0.200	-0.177
			(0.155)	(0.166)			(0.123)	(0.123)
Constant	1.739***	2.107***	1.516***	1.833***	5.402***	5.638***	5.252***	5.506***
	(0.305)	(0.231)	(0.312)	(0.238)	(0.245)	(0.204)	(0.253)	(0.208)

Table 2.10: Continued

	1	2	3	4	5	6	7	8
	Number of sales	Number of sales	Number of sales	Number of sales	Net-revenues	Net-revenues	Net-revenues	Net-revenues
	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions			Interactions	Interactions
Fixed effects								
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71927	62823	71927	62823	71927	62823	71927	62823
Individuals	248	216	248	216	248	216	248	216
R^2	0.193	0.210	0.197	0.215	0.311	0.331	0.313	0.332

The table presents OLS estimates. Log number of sales and log net-revenues are conditional on nonzero sales. Reciprocity is measured by the slope coefficient of a regression of back transfers on investments. Robust standard errors clustered on individual levels are in parentheses. In specifications 1 to 4, we estimate the effect of the average RIR on the log number of sales, while in specifications 5 to 8, we estimate the effect of the average RIR on log net-revenues. In specifications 2, 4, 6 and 8, observations from individuals with an average RIR > 2 are excluded (n=28). Furthermore, for applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2. This exclusion ensures that there are at least two individuals per store. Observations are on individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.11: Probability of making a sale - RC with intercept

	1	2	3	4
	All	Excl. RIR>2	All	Excl. RIR>2
			Interactions	Interactions
RC	-0.095	0.046	0.024	-0.114
	(0.133)	(0.136)	(0.170)	(0.265)
Medium consulting	-0.174	0.188	0.146	-0.039
	(0.200)	(0.209)	(0.338)	(0.454)
High consulting	0.590*	0.803**	0.604	0.437
	(0.313)	(0.368)	(0.646)	(0.689)
Full-time	-0.055	-0.218	-0.038	-0.225
	(0.180)	(0.209)	(0.181)	(0.213)
Tenure in months	0.002	0.001	0.002	0.002
	(0.001)	(0.002)	(0.001)	(0.001)
Male	-0.038	0.108	-0.010	0.151
	(0.223)	(0.238)	(0.230)	(0.259)
Medium consulting x RC			-0.229	0.181
			(0.194)	(0.338)
High consulting x RC			-0.009	0.294
			(0.356)	(0.442)
Constant	1.131*	0.367	0.967	0.564
	(0.591)	(0.624)	(0.615)	(0.705)
Fixed effects				
Month	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes
Observations	80236	69395	80236	69395
Individuals	247	215	247	285
Pseudo R^2	0.294	0.342	0.296	0.343
Wald Chi^2	1997.76	3508.76	1434.27	4059.03

The table presents probit estimates. Reciprocity is measured by the slope coefficient of a regression of back transfers on investments. Robust standard errors clustered on individual levels are in parentheses. In specifications 2 and 4, observations from individuals with an average RIR $>$ 2 are excluded (n=28). Furthermore, for applying store fixed effects, we have to exclude four additional individuals with an avg. RIR $<=$ 2. This exclusion ensures that there are at least two individuals per store. Observations of one individual are dropped, because of perfectly predicted success. Observations are on individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.12: Robustness checks - log net-revenues per sale

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark	Trusting	Demogr.	Pref.	Big 5	All
							Interactions	Interactions	Interactions	Interactions	Interactions	Interactions
							excl. RIR>2	excl. RIR>2	excl. RIR>2	excl. RIR>2	excl. RIR>2	excl. RIR>2
Avg. RIR	0.061**	0.065**	0.056**	0.058**	0.051*	0.055*	-0.025	-0.015	-0.040	-0.033	-0.059	-0.063
	(0.028)	(0.033)	(0.028)	(0.028)	(0.028)	(0.031)	(0.069)	(0.072)	(0.073)	(0.068)	(0.070)	(0.073)
Return if investment = 0	0.002	0.002	0.004	0.003	0.003	0.004	0.004	0.003	0.007	0.005	0.006	0.008
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Medium consulting	0.390***	0.375***	0.408***	0.394***	0.405***	0.411***	0.305**	0.327**	0.300*	0.303**	0.218	0.255*
	(0.048)	(0.047)	(0.049)	(0.048)	(0.050)	(0.050)	(0.152)	(0.153)	(0.155)	(0.152)	(0.144)	(0.146)
High consulting	0.756***	0.750***	0.784***	0.760***	0.758***	0.782***	0.447**	0.451**	0.427**	0.444**	0.404**	0.399**
	(0.065)	(0.063)	(0.065)	(0.066)	(0.064)	(0.064)	(0.178)	(0.179)	(0.180)	(0.178)	(0.168)	(0.173)
Full-time	0.130***	0.120**	0.111**	0.131***	0.129***	0.115**	0.103**	0.092*	0.077	0.108**	0.105**	0.082
	(0.048)	(0.048)	(0.049)	(0.048)	(0.048)	(0.048)	(0.052)	(0.053)	(0.055)	(0.052)	(0.053)	(0.057)
Tenure in months	0.001	0.001	0.000	0.001	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.193***	0.214***	0.059	0.193***	0.190***	0.105	0.198***	0.220***	0.061	0.203***	0.189***	0.100
	(0.060)	(0.060)	(0.075)	(0.060)	(0.060)	(0.075)	(0.065)	(0.068)	(0.075)	(0.066)	(0.064)	(0.079)
Trustor=6		-0.026				-0.017		-0.010				0.004
		(0.065)				(0.071)		(0.069)				(0.073)
Trustor=12		0.074				0.057		0.082				0.049
		(0.075)				(0.079)		(0.085)				(0.086)
Trustor=18		-0.085				-0.074		-0.074				-0.023
		(0.079)				(0.086)		(0.090)				(0.099)
Body height in cm			0.011***			0.008**			0.012***			0.010***
			(0.003)			(0.003)			(0.003)			(0.004)
Higher education			-0.059			0.004			-0.157			-0.102
			(0.097)			(0.095)			(0.107)			(0.107)
Other education			-0.097			-0.053			-0.130*			-0.103
			(0.078)			(0.075)			(0.070)			(0.071)
Risk preferences				-0.003		-0.004				-0.011		-0.012
				(0.009)		(0.010)				(0.010)		(0.011)
Time preferences				0.007		-0.001				0.010		-0.000
				(0.008)		(0.008)				(0.008)		(0.009)
Conscientiousness					0.007	0.011					-0.012	-0.006
					(0.034)	(0.034)					(0.033)	(0.035)
Extraversion					-0.015	-0.015					-0.018	-0.015
					(0.022)	(0.023)					(0.025)	(0.027)
Agreeableness					-0.031	-0.020					-0.032	-0.024
					(0.023)	(0.024)					(0.025)	(0.027)
Openness					0.029	0.024					0.025	0.022
					(0.023)	(0.023)					(0.024)	(0.023)
Neuroticism					-0.046***	-0.038*					-0.055***	-0.047**

TABLE 2.12: Continued

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark Interactions excl. RIR>2	Trusting Interactions excl. RIR>2	Demogr. Interactions excl. RIR>2	Pref. Interactions excl. RIR>2	Big 5 Interactions excl. RIR>2	All Interactions excl. RIR>2
					(0.018)	(0.020)					(0.019)	(0.022)
Medium consulting x avg. RIR							0.081	0.053	0.100	0.091	0.147	0.133
							(0.096)	(0.099)	(0.102)	(0.097)	(0.096)	(0.102)
High consulting x avg. RIR							0.259^{**}	0.248^{**}	0.287^{**}	0.270^{**}	0.284^{**}	0.303^{**}
							(0.115)	(0.117)	(0.118)	(0.115)	(0.111)	(0.117)
Constant	3.621^{***}	3.612^{***}	1.802^{***}	3.580^{***}	3.869^{***}	2.420^{***}	3.735^{***}	3.684^{***}	1.663^{***}	3.705^{***}	4.205^{***}	2.442^{***}
	(0.114)	(0.138)	(0.566)	(0.128)	(0.280)	(0.685)	(0.162)	(0.170)	(0.588)	(0.175)	(0.310)	(0.697)
Fixed effects												
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71927	71927	70513	71927	70857	70513	62823	62823	61409	62823	62823	61409
Individuals	248	248	241	248	244	241	216	216	209	216	216	209
R^2	0.359	0.362	0.365	0.360	0.363	0.369	0.373	0.375	0.381	0.374	0.377	0.385

The table presents OLS estimates. Log net-revenues per sale are conditional on nonzero sales. Robust standard errors clustered on individual levels are in parentheses. In specifications 6 to 10, observations from individuals with an average RIR > 2 are excluded (n=28). Furthermore, five individuals did not specify their body height and four individuals did not answer at least one item of the Big 5 inventory. For applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2 in specifications 6 to 10 and two individuals who indicated their body height in specifications 3, 5, 8 and 10. This exclusion ensures that there are at least two individuals per store. Observations are on individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 2.13: Robustness checks - log number of sales

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark Interactions excl. RIR>2	Trusting Interactions excl. RIR>2	Demogr. Interactions excl. RIR>2	Pref. Interactions excl. RIR>2	Big 5 Interactions excl. RIR>2	All Interactions excl. RIR>2
Avg. RIR	-0.128**	-0.161**	-0.124**	-0.118**	-0.116*	-0.160**	0.074	0.072	0.081	0.082	0.112	0.078
	(0.060)	(0.068)	(0.060)	(0.058)	(0.060)	(0.067)	(0.098)	(0.101)	(0.104)	(0.100)	(0.100)	(0.108)
Return if investment = 0	-0.004	-0.001	-0.004	-0.006	-0.006	-0.004	-0.004	0.003	-0.005	-0.006	-0.005	-0.002
	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)	(0.011)	(0.010)	(0.011)	(0.010)	(0.010)	(0.010)	(0.012)
Medium consulting	0.170*	0.186*	0.186*	0.167*	0.180*	0.245**	0.317	0.285	0.343	0.360	0.418*	0.440**
	(0.099)	(0.099)	(0.099)	(0.097)	(0.102)	(0.103)	(0.221)	(0.222)	(0.226)	(0.227)	(0.215)	(0.222)
High consulting	0.150	0.161	0.166	0.165	0.150	0.234*	0.796***	0.855***	0.843***	0.801***	0.848***	0.942***
	(0.123)	(0.123)	(0.121)	(0.120)	(0.126)	(0.124)	(0.272)	(0.266)	(0.283)	(0.267)	(0.267)	(0.264)
Full-time	-0.105	-0.115	-0.080	-0.094	-0.077	-0.061	-0.076	-0.080	-0.029	-0.075	-0.069	-0.018
	(0.092)	(0.086)	(0.094)	(0.092)	(0.094)	(0.089)	(0.100)	(0.092)	(0.106)	(0.099)	(0.101)	(0.100)
Tenure in months	0.001**	0.001**	0.001**	0.001*	0.001**	0.001*	0.001	0.001	0.001*	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Male	0.006	-0.028	-0.055	0.017	-0.002	-0.089	0.012	0.001	-0.043	0.031	0.017	-0.038
	(0.110)	(0.114)	(0.132)	(0.109)	(0.107)	(0.136)	(0.109)	(0.114)	(0.125)	(0.108)	(0.106)	(0.126)
Trustor=6		-0.044				0.016		-0.178				-0.127
		(0.107)				(0.111)		(0.108)				(0.111)
Trustor=12		-0.037				0.030		-0.118				-0.054
		(0.121)				(0.127)		(0.129)				(0.131)
Trustor=18		0.178				0.324*		0.006				0.085
		(0.165)				(0.182)		(0.174)				(0.188)
Body height in cm			0.004			0.004			0.002			0.002
			(0.006)			(0.006)			(0.006)			(0.005)
Higher education			-0.015			-0.018			0.140			0.135
			(0.204)			(0.211)			(0.236)			(0.256)
Other education			0.145			0.046			0.236*			0.215
			(0.153)			(0.173)			(0.138)			(0.156)
Risk preferences				-0.027		-0.048**				-0.020		-0.033*
				(0.017)		(0.020)				(0.018)		(0.019)
Time preferences				-0.016		-0.008				-0.017		-0.004
				(0.014)		(0.014)				(0.015)		(0.013)
Conscientiousness					0.068	0.079					0.083	0.085
					(0.056)	(0.053)					(0.053)	(0.053)
Extraversion					0.040	0.054					0.023	0.034
					(0.041)	(0.042)					(0.044)	(0.047)
Agreeableness					-0.054	-0.044					-0.033	-0.018
					(0.044)	(0.043)					(0.045)	(0.045)
Openness					-0.026	-0.001					-0.001	0.015
					(0.034)	(0.038)					(0.033)	(0.036)
Neuroticism					0.011	0.004					0.055	0.040

TABLE 2.13: Continued

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark	Trusting	Demogr.	Pref.	Big 5	All
							Interactions	Interactions	Interactions	Interactions	Interactions	Interactions
							excl. RIR>2	excl. RIR>2	excl. RIR>2	excl. RIR>2	excl. RIR>2	excl. RIR>2
					(0.038)	(0.035)					(0.038)	(0.039)
Medium consulting x avg. RIR							-0.120	-0.090	-0.119	-0.148	-0.174	-0.138
							(0.142)	(0.143)	(0.145)	(0.147)	(0.144)	(0.151)
High consulting x avg. RIR							-0.511***	-0.547***	-0.525***	-0.507***	-0.538***	-0.547***
							(0.193)	(0.188)	(0.198)	(0.191)	(0.190)	(0.183)
Constant	1.875***	1.962***	1.177	2.052***	1.604***	0.854	1.733***	1.842***	1.306	1.924***	1.060**	0.672
	(0.232)	(0.242)	(1.013)	(0.274)	(0.465)	(1.040)	(0.223)	(0.232)	(0.942)	(0.244)	(0.427)	(0.972)
Fixed effects												
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71927	71927	70513	71927	70857	70513	62823	62823	61409	62823	62823	61409
Individuals	248	248	241	248	244	241	216	216	209	216	216	209
R^2	0.195	0.198	0.194	0.198	0.194	0.204	0.218	0.221	0.218	0.220	0.221	0.225

The table presents OLS estimates. Log number of sales are conditional on nonzero sales. Robust standard errors clustered on individual levels are in parentheses. In specifications 6 to 10, observations from individuals with an average RIR > 2 are excluded (n=28). Furthermore, five individuals did not specify their body height and four individuals did not answer at least one item of the Big 5 inventory. For applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2 in specifications 6 to 10 and two individuals who indicated their body height in specifications 3, 5, 8 and 10. This exclusion ensures that there are at least two individuals per store. Observations are on individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.14: Robustness checks - log net-revenues

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark Interactions excl. RIR>2	Trusting Interactions excl. RIR>2	Demogr. Interactions excl. RIR>2	Pref. Interactions excl. RIR>2	Big 5 Interactions excl. RIR>2	All Interactions excl. RIR>2
Avg. RIR	-0.067	-0.096*	-0.068	-0.060	-0.065	-0.106*	0.049	0.057	0.041	0.048	0.053	0.015
	(0.051)	(0.057)	(0.050)	(0.049)	(0.050)	(0.055)	(0.093)	(0.089)	(0.093)	(0.092)	(0.096)	(0.092)
Return if investment = 0	-0.002	0.001	-0.000	-0.003	-0.003	-0.001	0.000	0.006	0.003	-0.001	0.002	0.006
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)
Medium consulting	0.560***	0.561***	0.594***	0.561***	0.585***	0.656***	0.622***	0.613***	0.643***	0.663***	0.636***	0.695***
	(0.080)	(0.081)	(0.081)	(0.078)	(0.084)	(0.081)	(0.185)	(0.176)	(0.181)	(0.185)	(0.190)	(0.177)
High consulting	0.907***	0.911***	0.950***	0.926***	0.907***	1.017***	1.243***	1.306***	1.270***	1.246***	1.252***	1.340***
	(0.101)	(0.101)	(0.098)	(0.100)	(0.102)	(0.098)	(0.214)	(0.211)	(0.212)	(0.207)	(0.213)	(0.197)
Full-time	0.025	0.004	0.031	0.037	0.052	0.054	0.028	0.011	0.048	0.033	0.036	0.065
	(0.079)	(0.075)	(0.078)	(0.078)	(0.081)	(0.074)	(0.087)	(0.083)	(0.088)	(0.085)	(0.089)	(0.084)
Tenure in months	0.002***	0.002***	0.002***	0.001***	0.001***	0.001***	0.001**	0.001***	0.001***	0.001**	0.001**	0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.198**	0.187**	0.004	0.211**	0.188**	0.016	0.210**	0.221**	0.018	0.234***	0.205**	0.062
	(0.088)	(0.092)	(0.108)	(0.087)	(0.086)	(0.109)	(0.090)	(0.092)	(0.102)	(0.088)	(0.089)	(0.102)
Trustor=6		-0.070				-0.001		-0.187**				-0.123
		(0.080)				(0.078)		(0.080)				(0.077)
Trustor=12		0.036				0.087		-0.037				-0.005
		(0.085)				(0.085)		(0.087)				(0.084)
Trustor=18		0.093				0.250		-0.068				0.062
		(0.152)				(0.153)		(0.157)				(0.156)
Body height in cm			0.015***			0.012***			0.015***			0.012***
			(0.005)			(0.004)			(0.004)			(0.004)
Higher education			-0.074			-0.014			-0.016			0.033
			(0.148)			(0.160)			(0.180)			(0.201)
Other education			0.048			-0.007			0.106			0.112
			(0.179)			(0.193)			(0.155)			(0.164)
Risk preferences				-0.030**		-0.052***				-0.031**		-0.044***
				(0.014)		(0.017)				(0.014)		(0.017)
Time preferences				-0.009		-0.009				-0.007		-0.005
				(0.013)		(0.012)				(0.013)		(0.011)
Conscientiousness					0.076*	0.090**					0.071	0.080*
					(0.043)	(0.040)					(0.046)	(0.044)
Extraversion					0.025	0.039					0.006	0.019
					(0.032)	(0.032)					(0.035)	(0.036)
Agreeableness					-0.085**	-0.064*					-0.065*	-0.043
					(0.037)	(0.035)					(0.038)	(0.035)
Openness					0.003	0.023					0.024	0.037
					(0.029)	(0.029)					(0.029)	(0.028)
Neuroticism					-0.035	-0.034					-0.000	-0.007

Table 2.14: Continued

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark Interactions excl. RIR>2	Trusting Interactions excl. RIR>2	Demogr. Interactions excl. RIR>2	Pref. Interactions excl. RIR>2	Big 5 Interactions excl. RIR>2	All Interactions excl. RIR>2
					(0.031)	(0.028)					(0.033)	(0.031)
Medium consulting x avg. RIR							-0.039	-0.037	-0.019	-0.057	-0.027	-0.005
							(0.123)	(0.119)	(0.121)	(0.126)	(0.128)	(0.125)
High consulting x avg. RIR							-0.252*	-0.299**	-0.238*	-0.237*	-0.254*	-0.244*
							(0.145)	(0.141)	(0.140)	(0.141)	(0.144)	(0.127)
Constant	5.496***	5.573***	2.978***	5.632***	5.473***	3.273***	5.468***	5.526***	2.969***	5.630***	5.265***	3.114***
	(0.205)	(0.181)	(0.826)	(0.234)	(0.385)	(0.775)	(0.207)	(0.187)	(0.767)	(0.216)	(0.386)	(0.740)
Fixed effects												
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71927	71927	70513	71927	70857	70513	62823	62823	61409	62823	62823	61409
Individuals	248	248	241	248	244	241	216	216	209	216	216	209
R^2	0.312	0.314	0.316	0.314	0.314	0.326	0.333	0.337	0.338	0.336	0.336	0.346

The table presents OLS estimates. Log net-revenues are conditional on nonzero sales. Robust standard errors clustered on individual levels are in parentheses. In specifications 6 to 10, observations from individuals with an average RIR > 2 are excluded (n=28). Furthermore, five individuals did not specify their body height and four individuals did not answer at least one item of the Big 5 inventory. For applying store fixed effects, we have to exclude four additional individuals with an avg. RIR <= 2 in specifications 6 to 10 and two individuals who indicated their body height in specifications 3, 5, 8 and 10. This exclusion ensures that there are at least two individuals per store. Observations are on individual daily level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.15: Robustness checks - probability of making a sale

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark Interactions excl. RIR>2	Trusting Interactions excl. RIR>2	Demogr. Interactions excl. RIR>2	Pref. Interactions excl. RIR>2	Big 5 Interactions excl. RIR>2	All Interactions excl. RIR>2
Avg. RIR	-0.128	-0.224*	-0.147	-0.133	-0.189	-0.292**	0.022	-0.163	-0.088	0.028	-0.135	-0.447*
	(0.133)	(0.124)	(0.133)	(0.132)	(0.139)	(0.117)	(0.272)	(0.227)	(0.255)	(0.272)	(0.269)	(0.263)
Return if investment = 0	0.011	0.003	0.015	0.010	0.022	0.018	0.043*	0.023	0.038	0.039	0.042*	0.008
	(0.022)	(0.023)	(0.022)	(0.023)	(0.022)	(0.023)	(0.024)	(0.026)	(0.026)	(0.024)	(0.025)	(0.026)
Medium consulting	-0.177	-0.150	-0.166	-0.156	-0.232	-0.199	0.172	0.135	0.103	0.311	-0.011	0.065
	(0.196)	(0.176)	(0.202)	(0.199)	(0.216)	(0.199)	(0.485)	(0.466)	(0.470)	(0.488)	(0.514)	(0.468)
High consulting	0.604*	0.557*	0.643**	0.651**	0.501	0.529*	0.696	0.572	0.626	0.903	0.839	0.714
	(0.318)	(0.314)	(0.306)	(0.318)	(0.324)	(0.307)	(0.840)	(0.839)	(0.676)	(0.827)	(0.767)	(0.629)
Full-time	-0.069	-0.106	-0.157	-0.046	-0.104	-0.195	-0.296	-0.367*	-0.337	-0.277	-0.367*	-0.382*
	(0.185)	(0.189)	(0.187)	(0.185)	(0.185)	(0.182)	(0.206)	(0.213)	(0.239)	(0.205)	(0.204)	(0.217)
Tenure in months	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Male	-0.048	-0.037	-0.344	-0.065	-0.050	-0.284	0.075	0.172	-0.233	0.042	-0.060	-0.400
	(0.223)	(0.222)	(0.304)	(0.222)	(0.231)	(0.309)	(0.239)	(0.241)	(0.266)	(0.239)	(0.239)	(0.290)
Trustor=6		0.474**				0.589***		0.247				0.369
		(0.212)				(0.218)		(0.224)				(0.238)
Trustor=12		0.457				0.609**		0.909***				1.151***
		(0.278)				(0.270)		(0.344)				(0.346)
Trustor=18		0.785*				0.706*		0.385				0.438
		(0.416)				(0.375)		(0.487)				(0.415)
Body height in cm			0.017			0.018			0.022			0.029**
			(0.014)			(0.013)			(0.014)			(0.015)
Higher education			0.088			0.180			-0.085			0.253
			(0.460)			(0.459)			(0.641)			(0.645)
Other education			-0.830			-0.844*			-0.274			-0.179
			(0.515)			(0.463)			(0.246)			(0.240)
Risk preferences				-0.038		-0.044				-0.053		-0.028
				(0.039)		(0.041)				(0.039)		(0.039)
Time preferences				-0.007		0.002				-0.016		0.026
				(0.037)		(0.037)				(0.037)		(0.040)
Conscientiousness					-0.217	-0.167					-0.424**	-0.394**
					(0.142)	(0.143)					(0.174)	(0.165)
Extraversion					-0.104	-0.110					0.003	0.004
					(0.088)	(0.077)					(0.083)	(0.083)
Agreeableness					-0.126	-0.112					-0.116	-0.072
					(0.115)	(0.113)					(0.115)	(0.123)
Openness					0.066	0.086					-0.027	-0.096
					(0.097)	(0.087)					(0.108)	(0.095)
Neuroticism					-0.074	-0.067					0.013	0.011

Table 2.14: Continued

	1	2	3	4	5	6	7	8	9	10	11	12
	Benchmark	Trusting	Demogr.	Pref.	Big 5	All	Benchmark Interactions excl. RIR>2	Trusting Interactions excl. RIR>2	Demogr. Interactions excl. RIR>2	Pref. Interactions excl. RIR>2	Big 5 Interactions excl. RIR>2	All Interactions excl. RIR>2
					(0.081)	(0.079)					(0.085)	(0.094)
Medium consulting x avg. RIR							0.012	0.004	0.122	-0.048	0.155	0.177
							(0.363)	(0.327)	(0.356)	(0.363)	(0.374)	(0.339)
High consulting x avg. RIR							0.100	0.085	0.229	0.027	0.047	0.204
							(0.515)	(0.487)	(0.443)	(0.498)	(0.467)	(0.419)
Constant	1.210^{*}	1.078	-1.533	1.379^{**}	3.837^{***}	0.247	0.435	0.245	-3.216	0.689	3.776^{***}	-1.814
	(0.628)	(0.666)	(2.461)	(0.658)	(1.297)	(2.679)	(0.705)	(0.644)	(2.636)	(0.723)	(1.372)	(2.984)
Fixed effects												
Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weekday	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Promotion day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Store	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	80236	80236	78638	80236	78991	78638	69395	69395	67797	69395	68150	67797
Individuals	247	247	240	247	243	240	215	215	208	215	211	208
Pseudo R^2	0.295	0.307	0.312	0.297	0.314	0.331	0.346	0.361	0.357	0.349	0.371	0.381
Wald Chi^2	1981.73	1672.04	2370.80	1342.06	1983.52	2786.41	9028.31	7919.01	5041.51	6088.97	53791.50	3981.86

The table presents probit estimates. Robust standard errors clustered on individual levels are in parentheses. In specifications 6 to 10, observations from individuals with an average RIR > 2 are excluded (n=28). Furthermore, five individuals did not specify their body height and four individuals did not answer at least one item of the Big 5 inventory. For applying store fixed effects, we have to exclude four additional individuals with an avg. RIR $<= 2$ in specifications 6 to 10 and two individuals who indicated their body height in specifications 3, 5, 8 and 10. This exclusion ensures that there are at least two individuals per store. Observations are on individual daily level. Observations of one individual are dropped, because of perfectly predicted success. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3 Essay 2: Choking under time pressure: The influence of deadline-dependent bonus and malus incentive schemes on performance

Andrea Essl, Stefanie Jaussi *

This is the working paper version of the article that is published in the Journal of Economic Behavior & Organization (http://dx.doi.org/10.1016/j.jebo.2016.11.001).

Abstract

Many economically relevant activities are executed under notable time pressure. Incentivized deadlines are often the reason people feel pressured. In such an environment, personal characteristics can be important drivers of performance. By means of a laboratory experiment, we examine the predictive power of individual loss aversion on performance under deadline-dependent incentives. Participants worked on a real effort task under two payoff-equivalent contracts, framed in bonus and malus terms. Results show lower performance for individuals with a high level of loss aversion when working under a malus contract. These persons gave fewer correct answers and needed more time to reply than other individuals. Choking can explain this observed behavior.

Keywords: time pressure, deadline-dependent incentive schemes, loss aversion, choking under pressure

*Andrea Essl (andrea.essl@iop.unibe.ch) and Stefanie Jaussi (stefanie.jaussi@iop.unibe.ch) are from the Institute of Organization and Human Resource Management, University of Bern. We are grateful to Sonja Erni and Simon Roth for providing excellent research assistance. We also thank Frauke von Bieberstein, Nils Kraizcy, Claudia Vogel and participants at EBES (Istanbul) and the internal seminar at the University of Bern for helpful comments. All errors are ours.

3.1 Introduction

One of the most important questions for organizations is how to make incentives most effective. An ongoing discussion concerns the choice between the carrot and the stick. Does one want to offer rewards for excellent performance or rather punish results that do not meet expectations? Standard economic theory predicts no difference between positively and negatively framed contracts that offer economically equivalent incentives. Yet considerable evidence from laboratory and field experiments reveals that malus contracts are more effective than bonus contracts (Hannan et al. 2005, Church et al. 2008, Brooks et al. 2012, Hossain and List 2012, Fryer et al. 2013). Prospect theory offers an explanation of this effect. As "losses loom larger than gains" (Kahneman and Tversky 1979), people invest more effort to prevent a possible loss than to achieve a bonus of the same value. The prevailing dominance of bonus contracts in practice is thus hard to explain.

In some situations, incentives do not have the desired influence on performance. Counterproductive effects can arise when incentives cause high pressure and result in worse instead of improved performance. The phenomenon whereby high pressure leads to a performance decline is called *choking under pressure* (Baumeister 1984). Besides performance-dependent incentives, competitive environment, the presence of an audience, and relevance of the ego (Baumeister and Showers 1986), the literature acknowledges time pressure as a stressor variable (Cavanaugh et al. 2000). Time is money, and therefore, deadlines are a major challenge in the work environment. To promote high work speed, incentives are linked to goal achievement within a given timeline. A prime example is the construction industry, where malus and bonus incentives are often linked to a deadline. Consequently, an important question for companies is how to structure incentive schemes in a time-pressured environment. The aim of our study is to address this issue. In particular, we investigate whether the level of individuals' loss aversion affects performance under deadline-dependent bonus and malus incentive schemes. This provides insights into the potential hidden costs of malus contracts. As people put greater value on losses than on gains, maluses consequently represent higher incentives. Thus, we hypothesize that for individuals with a high degree of loss aversion a malus incentive becomes a more important pressure variable which can then, under time pressure, lead to choking.

Despite the high relevance of the topic, literature examining the consequences of time-dependent incentives is limited. Payne et al. (1996) analyzed risky choices where time pressure was implemented by increasing costs for delayed decisions. They showed that a time-dependent incentive scheme led to lower payoffs. Kocher and Sutter (2006) examined the trade-off between quality and time in strategic decision-making and observed that time pressure decreases quality. Interestingly, they found that the negative consequences of time pressure can be counterbalanced by a time-dependent incentive scheme. By multiplying payoffs by a time-dependent factor, they incorporated both bonus and malus elements in the incentive scheme used in their study. Faster decisions were rewarded with a factor larger than one and slower decisions were punished with a factor smaller than one. Kocher and Sutter (2006) suggested that the use of time-dependent incentives can be a valuable option in a time-pressured environment, but it remains unclear to what extent and under which circumstances the bonus or the malus elements are responsible for the positive influence. In our study, we take up these concerns by separately examining the impact of bonus and malus contracts on performance under time pressure.

In our experiment, participants completed a real effort task wherein both time and quality determined performance. The incentive schemes contingent on a deadline were structured as follows. In both treatments, participants were rewarded for fulfilling the task. Additionally, in the bonus treatment, participants who entered the answer within the first 10 seconds received a bonus and in the malus treatment, participants who needed more than 10 seconds got a malus. Incentives were payoff-equivalent across treatments. Only the framing in terms of the bonus or malus contract varied. A loss aversion test (Fehr and Goette 2007, Gächter et al. 2010, Abeler et al. 2011) enable us to identify the predictive power of the individual's degree of loss aversion regarding performance. The results show that when working under a deadline-dependent malus incentive scheme, individuals with a high level of loss aversion reported worse performance, whereas individuals with a very low degree of loss aversion increased their performance. Performance differences are driven by both response time and correct answers. With respect to the response time, it is interesting that individuals with a high level of loss aversion were less able to answer the task within 10 seconds and thus significantly suffered more malus payments. Altogether, our results complement the existing literature and emphasize the advantage of including loss aversion in the design of incentive schemes in a time-pressured environment. The main contribution of this paper is the finding that how loss aversion influences individual performance depends on the received economic incentives. Thus, to achieve the desired effects of incentives, companies should not only consider different incentive elements but also behavioral biases when writing contracts.

The rest of the paper is organized as follows. In Section 3.2, we review the relevant literature. The experimental design is introduced in Section 3.3, and Section 3.4 presents the hypotheses. The results of the experiment are described in Section 3.5 and discussed in Section 3.6, which concludes the paper.

3.2 Literature review

3.2.1 Framing effects and individual loss aversion

A broad range of empirical work has examined the impact of contract frames (Luft 1994, Hannan et al. 2005, Church et al. 2008, Hossain and List 2012, Fryer et al. 2013). For example, Luft (1994) studied contract choice and showed that individuals prefer contracts formulated in bonus terms to malus contracts. By extension, Hannan et al. (2005) found that even though individuals prefer bonus contracts, the effect of loss aversion leads to higher effort choice under malus contracts. Furthermore, Church et al. (2008) added to the literature by showing that not only chosen effort but also real effort provision increases under a malus contract. Recently, Hossain and List (2012) and Fryer et al. (2013) tested framing manipulations in the field. Hossain and List (2012) observed significantly higher team productivity under payments framed in malus terms in a high-tech manufacturing company in China. Fryer et al. (2013) examined the influence of framed contracts for teachers on students' performance. When teachers were paid according to a loss contract, students improved their math exam scores significantly more than those instructed by a teacher with a gain contract. Recent, conflicting results from experimental studies raise questions about these findings (Imas et al. 2015, Quidt 2014, Hilken et al. 2013). Imas et al. (2015) and Quidt (2014) suggested a higher acceptance rate for malus contracts, whereas Hilken et al. (2013) found higher effort levels for bonus contracts than for malus or combined incentive schemes. Thus, a more detailed analysis of framing effects under different conditions is clearly necessary.

One important aspect is the interaction of situational elements and individual characteristics. Several studies have shown that individual loss aversion preferences can be important predictors of behavior in the workplace (Fehr and Goette 2007, Fehr et al. 2008, Brink and Rankin 2013). Fehr and Goette (2007) conducted a field experiment with bike messengers and showed that those with high levels of individual loss aversion reacted to a temporarily higher wage with less effort per shift. Similarly, Fehr et al. (2008) examined the relation between loss aversion and effort. In their experiment, participants were paid to enter data into a computer program while facing random delays. When working under a piece rate instead of a fixed wage, individuals with high loss averse preferences responded to the delay by investing more effort. Brink and Rankin (2013) found that contract preferences under differently framed incentives are largely influenced by individual loss aversion.

3.2.2 Choking under time pressure

Two types of psychological theories can explain the phenomena *choking under pressure*. First, distraction theory postulates that pressure causes a focus shift away from the actual task towards worrisome thoughts. Second, explicit monitoring theory indicates that pressure increases anxiety about failure and that it is the explicit focus on the task that disrupts proceduralized performance (Beilock and Carr 2001, DeCaro et al. 2011, Sanders and Walia 2012). Thus, the two theories suggest opposing mechanisms, both resulting in skill failure. Distraction theory gives a reason for a shift of attention away from the task, and explicit monitoring theory describes an excessively strong shift towards the task (DeCaro et al. 2011). Pressure variables which lead to a performance decrease can be manifold. Evidence has shown that high expectations of the audience (Dohmen 2008), very high monetary incentives (Ariely et al. 2009), and competition (Smith 2013) lead to choking. Moreover, individual characteristics can be important determinants of whether choking occurs in a given situation or not. Deffenbacher (1978) showed in an experiment that individuals with high levels of test anxiety reported worse performance under a high stress condition. Moreover, Baumeister et al. (1993) found that for people with a high level of self-esteem, threats of the ego led to smaller rewards, whereas they performed better without ego stimulation. Similarly, time pressure can amplify other sources of pressure and lead to choking (Shurchkov 2012, Bracha and Fershtman 2013). Shurchkov (2012) showed that women underperformed men only in a competitive stereotypical task when under time pressure. Furthermore, Bracha and Fershtman (2013) observed a performance decrease in the form of lower success rates when people were under time pressure in tournaments but no mere effect of time pressure on those working on a piece rate basis. These findings show that depending on the environment, time pressure can lead to significant performance decreases.

As mentioned in Section 3.1, only a small stream of research has examined the performance consequences of time-dependent incentive schemes (Payne et al. 1996, Kocher and Sutter 2006). Payne et al. (1996) implemented time-dependent incentives in risky choices, whereas Kocher and Sutter (2006) used them in a strategic setting. Moreover, both multiplied the payoffs with a time-dependent factor. How deadline-dependent bonus and malus contracts influence individual performance in a risk-free, non-strategic setting remains therefore an open question.

3.3 Experimental design and procedure

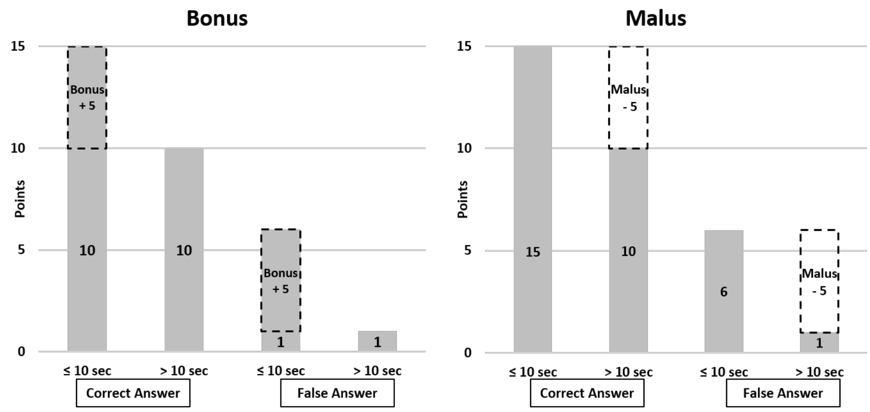
3.3.1 Experimental design

We used a real effort task to examine how deadline-dependent incentive schemes influence the quality and speed of performance. Participants had to count the number of zeros in tables consisting of zeros and ones (Abeler et al. 2011). Each table had four rows of different length, randomly filled with zeros and ones. The number of zeros varied between 20 and 25.[1] This task offers several advantages. It requires no special knowledge, the correctness and speed of task performance are easily measurable, and learning possibilities are trivial. Furthermore, because of its apparent artificiality any tendency towards intrinsic or reciprocal motivation can be minimized. In the main stage of the experiment, participants worked on 20 tasks.

The experiment consisted of two treatments. In one treatment, participants could receive a bonus for entering the answer before a given deadline. In the other treatment, participants obtained a malus when exceeding the deadline. The two performance measures were quality and speed. We designed economically equivalent bonus and malus incentive schemes that rewarded both of these measures. Figure 3.1 provides a summary of the incentive schemes. Participants in the bonus treatment earned a piece rate of 10 points for correct and one point for false answers. In addition, if they were able to provide an answer in the first 10 seconds, they received a bonus of five points independently of the correctness of the answer. In the malus treatment, the piece rate was 15 points for correct answers and six points for false responses. In contrast to the bonus treatment, individuals received a malus of five points if they failed to give an answer within the first 10 seconds. Thus, in both treatments, for correct answers participants could earn 15 points when they gave the answer within 10 seconds and 10 points when they exceeded the 10-second threshold. For false answers, participants received six points when they entered an answer within the first 10 seconds and one point when they provided the answer after 10 seconds. We introduced earnings for false answers so the bonus and the malus components were based strictly on speed. Furthermore, our incentive schemes resembled those often used in practice which ensure payments for effort and provide additional incentives for extraordinary performance. Additionally, without earnings for false answers, participants in the malus treatment would incur a loss of five points for giving a wrong answer after 10 seconds.

[1] Experimental instructions including a screenshot of the task can be found in Appendix A.

Figure 3.1: Deadline-dependent bonus and malus incentive schemes



After completing a table, participants received direct feedback on whether the answer was correct or false and depending on the treatment whether they received a bonus or a malus. Then the next table was shown. We set the maximum response time for each table to 20 seconds. If no entry was made within 20 seconds, a message indicated that no points were earned for this table. A pilot session had shown that the average time for answering was about 13 seconds. Therefore the time limit of 20 seconds was chosen to induce negligible time pressure on participants.[2] After completing the 20 tables, participants were informed about the total number of correct and false answers, the total number of unsolved tables, and the number of points earned. At the end of the experiment, the accumulated points were translated into monetary outcomes at an exchange rate of 10 points = 0.4 CHF (at the time of the experiment, 1 CHF =1.05 USD).

Before and after the main stage, we introduced two identical stages with five tables each. In these stages, there was no deadline-dependent incentive scheme and participants earned the same piece rate for correct and false answers as in the main stage. The first stage provided a measurement of individuals' baseline performance. In addition, it allowed participants to get used to the task. The stage after the main stage was introduced to analyze possible fatigue effects and to capture participants' reaction to the removal of the deadline-dependent incentive scheme. In the last stage of the experiment, participants took a loss aversion test (Fehr and Goette 2007, Gächter et al. 2010, Abeler et al. 2011) and filled out a short questionnaire on demographics. In the loss aversion test, participants had to make six choices of whether or not to play a lottery.[3] In each lottery the winning price was fixed at 6 CHF and the losing price varied from -2 to -7 CHF.

[2]In all treatments and out of 1920 tables only nine were not completed within 20 seconds. In the main stage, two tables were not completed within 20 seconds.
[3]The loss aversion test can be found in Appendix A.

There was a 50/50 chance getting 6 CHF or receiving the losing price. At the end of the experiment, one lottery was randomly chosen and paid out (Cubitt et al. 1998).

3.3.2 Procedure

In total, 69 students from various disciplines of the University of Bern took part in the experiment. Five participants were excluded from data analysis because their answers in the loss aversion test were inconsistent.[4] Of the remaining 64 participants, 34 were in the bonus treatment and 30 in the malus treatment. The experiment was conducted in the AareLab of the University of Bern in May and June 2013. All participants were randomly assigned to the treatments and treatments were randomized over morning and afternoon sessions as well as over weekdays. Subjects were recruited via ORSEE (Greiner 2004). The experiment was computerized using z-Tree (Fischbacher 2007). The participants' average age was 23 years and 53% were female.

We paid special attention to avoiding potential confounds like peer pressure (Falk and Ichino 2006) and a desire for conformity (Bernheim 1994). In particular, we wanted to make sure that participants were not influenced by the speed of others who could potentially be heard when entering their answers on the computer keyboard. Thus, we decided to have only one person per session. Sessions lasted between 25 and 30 minutes. Earnings averaged 21 CHF, including a show-up fee of 6 CHF. Participants received written instructions for each stage separately and were asked to solve some control questions. At the end of the experiment, they were paid directly in CHF.

3.4 Hypotheses

We consider a situation where monetary incentives across the bonus and the malus treatment are the same, but are perceived differently given reference-dependent preferences. In particular, individuals with a high level of loss aversion have a higher incentive to avoid a malus than individuals with a low level of loss aversion. Therefore, people with high loss averse preferences are more prone to choking under pressure, which, in turn, decreases performance. Accordingly, we hypothesize:

Hypothesis 1: *On average, individuals with a high level of loss aversion working under a deadline-dependent malus incentive scheme earn fewer profit points than all other individuals.*

[4]These five participants switched more than once between rejecting and accepting the lottery and thus displayed non-monotonicity. Following Gächter et al. (2010) and Abeler et al. (2011), we decide to exclude them from the analyses.

In the context of the study, choking can emerge in different ways. Given that the incentive scheme is based on a deadline, choking may cause weaker performance in the time dimension. Therefore, individuals with a high degree of loss aversion in the malus treatment are more likely to answer fewer questions within the first 10 seconds and thus suffer more malus payments. As we analyze both the absolute response time and the number of avoided maluses, our findings reveal whether participants with a high level of loss aversion in the malus treatment need more time to answer but are still able to answer within the 10-second deadline or whether they are less able to meet this deadline and thus receive more malus payments. Only in the latter case does one find an impact on profit points, because of the monetary disadvantage of answering after 10 seconds. Therefore, we derive the following hypothesis:

Hypothesis 2: *On average, individuals with a high level of loss aversion working under a deadline-dependent malus incentive scheme need more time to respond and thus are less able to avoid malus payments than all other individuals.*

Next to the time dimension, there may be also a quality decrease. A decrease in quality could be explained by distraction theory. Our incentive scheme may create a distracting environment, which takes the attention from the actual task to the deadline-dependent incentive. Specifically, in the malus treatment people with a high level of loss aversion may struggle with worrisome thoughts about receiving a malus and thus the focus, which is normally strongly devoted to the task quality, now competes with worries about getting a malus payment. Therefore, we hypothesize:

Hypothesis 3: *On average, individuals with a high level of loss aversion working under a deadline-dependent malus incentive scheme give fewer correct answers than all other individuals.*

3.5 Results

We examine the speed and quality of task performance and the forces behind the influence of deadline-dependent incentives. The measures of interest are the accumulated profit points, the response time, the number of bonuses received or, respectively, the number of maluses avoided, and the number of correct answers. Table 3.1 reports descriptive statistics. It shows that participants in the bonus treatment accumulated more profit points than individuals in the malus treatment (p=0.205; Mann-Whitney test).[5] Furthermore, individuals in the bonus treatment had a lower average response time per task (p=0.581; Mann-Whitney test), which, in turn, led to a higher average number

[5]All statistical tests are two-sided.

of received bonuses than avoided maluses (p=0.257; Mann-Whitney test). In addition, Table 3.1 indicates that compared with the malus treatment, participants in the bonus treatment gave, on average, slightly more correct answers (p=0.315; Mann-Whitney test). However, according to the Mann-Whitney tests the differences are not significant.[6]

TABLE 3.1: Descriptive statistics

Treatment		Profit	Response time	Bonus received/ malus avoided	Correct answers
Malus (n=30)	Mean	251.63	8.87	15.77	17.00
	S.D.	51.59	1.69	5.72	3.09
Bonus (n=34)	Mean	265.26	8.28	17.29	17.65
	SD	37.99	1.73	4.00	2.92

The table shows means and standard deviations.

As we proposed in our hypotheses, there might be a relationship between a participant's degree of loss aversion and his or her performance under deadline-dependent incentive schemes. Thus, we analyze whether loss averse preferences have a predictive value for behavior in the bonus and malus treatments. The conducted loss aversion test (Fehr and Goette 2007, Gächter et al. 2010, Abeler et al. 2011) enables us to construct an individual measurement of loss aversion. In each out of six lotteries, participants could either win 6 CHF or receive the losing price with 50% probability. The losing price varied in integer values from -2 CHF in lottery number one to -7 CHF in lottery number six. Lotteries number one to five had non-negative expected values. In both treatments, more than 40% of participants rejected at least lottery number four with the losing price of -5 CHF. According to Rabin (2000), rejections of small-stake lotteries with positive expected value can serve as an indicator for the individual's degree of loss aversion. For bivariate analyses we classified participants who rejected lotteries with a losing price higher than -5 CHF (lotteries one to three) as individuals with a high level of loss aversion and participants who accepted lotteries even with a losing price equal to or smaller than -5 CHF as those with a low level of loss aversion (lotteries four to six). In the regression analysis, we use the number of rejected lotteries as a continuous measurement of loss aversion.

The remaining results are organized as follows. After we show the impact of loss averse preferences on the accumulated profit points under the different deadline-dependent incentive schemes, we analyze whether quality, speed or the combination of both drive possible performance differences between individuals with different degrees of loss aversion.

[6]Results for each single task are presented in Appendix B.

3.5.1 Profit points

Table 3.2 presents descriptive statistics for the accumulated profits of individuals with a high level of loss aversion and those with a low level of loss aversion. The maximum number of reachable profit points was set to 300. Profits of participants with a high level of loss aversion are significantly lower in the malus treatment than in the bonus treatment (p=0.049; Mann-Whitney test). In addition, participants displaying high loss averse preferences in the malus treatment earned significantly less than individuals with a low level of loss aversion in both treatments (for the bonus treatment p=0.027; for the malus treatment p=0.028; Mann-Whitney test). There is neither a significant treatment effect between individuals with a low degree of loss aversion (p=0.927; Mann-Whitney test) nor significant differences in the means of accumulated profits between people with different degrees of loss aversion in the bonus treatment (p=0.888; Mann-Whitney test). In addition, we do not find significant differences between individuals with a high degree of loss aversion in the bonus treatment and participants with a low degree of loss aversion in the malus treatment (p=0.889; Mann-Whitney test).

TABLE 3.2: Descriptive statistics of profit points

	Malus HLA (n=13)	Malus LLA (n=17)	Bonus HLA (n=14)	Bonus LLA (n=20)
Mean	225.23	271.82	267.50	263.70
S.D.	64.36	26.81	34.43	41.10

The table shows means and standard deviations for accumulated profits of individuals with a high level of loss aversion (HLA) and those with a low level of loss aversion (LLA).

The question of whether there is a treatment effect, when we consider individuals' degree of loss aversion, on profits is also addressed in Table 3.3. It presents the estimates of four specifications of a random effects model. The dependent variable is accumulated profit points. We apply robust standard errors clustered on an individual level. In specification 1, we regress the profit points on the malus treatment dummy variable and the individual's degree of loss aversion. The treatment dummy is equal to one for the malus treatment and zero for the bonus treatment. Contrary to the bivariate analyses where we distinguish between individuals with high and low degrees of loss aversion, we use the continuous measurement of loss aversion in the regression analyses. This measurement is represented by the number of rejected lotteries in the loss aversion test.

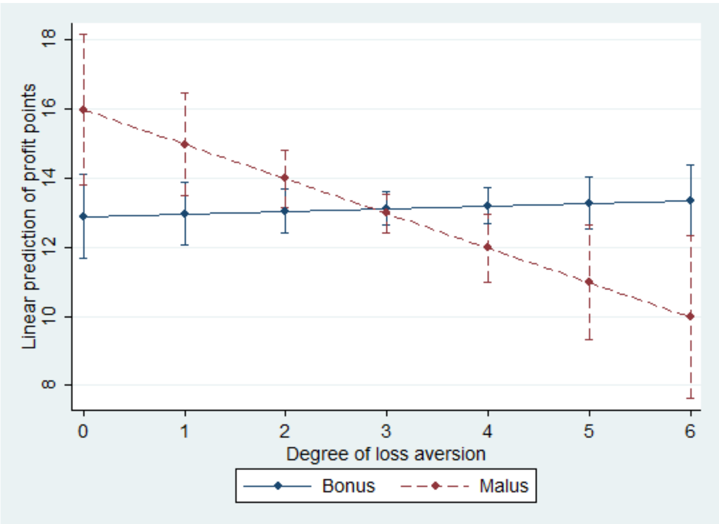
Table 3.3: Determinants of profit points

	1	2	3	4
Malus	-0.666	3.709***	3.086**	2.558**
	(0.530)	(1.336)	(1.250)	(1.279)
Loss aversion	-0.509**	0.066	0.076	0.044
	(0.256)	(0.250)	(0.171)	(0.157)
Malus x loss aversion		-1.345***	-1.077***	-0.935**
		(0.425)	(0.410)	(0.402)
Baseline performance			0.486***	0.501***
			(0.083)	(0.078)
Age				-0.071**
				(0.035)
Female				-0.706
				(0.460)
Years of study				-0.099
				(0.148)
Constant	14.909***	13.049***	9.786***	11.979***
	(0.917)	(0.923)	(0.904)	(1.366)
N	1280	1280	1280	1220
Individuals	64	64	64	61
R^2 *overall*	0.038	0.091	0.177	0.199
Chi^2	4.149	13.975	47.634	92.577
Sigmau	2.031	1.837	1.436	1.413
Rho	0.285	0.246	0.166	0.162

The table shows estimates from a random effects model. Robust standard errors clustered on individual levels are in parentheses. Sample sizes differ because three students did not answer the question on how many years they had studied. Loss aversion is measured by the number of rejected lotteries in the loss aversion test. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In specification 2, we add the interaction between the malus treatment and the individual's degree of loss aversion. In specification 3, we include the baseline performance and in specification 4, we additionally control for age, gender, and years of study. Baseline performance is measured in the first stage, where no deadline-dependent incentive was in place, by multiplying average correct answers with the remaining time after all answers had been given. Specification 1 shows a significant negative influence of a higher degree of loss aversion on profit points. Allowing for an interaction between the malus treatment and the individual's degree of loss aversion, we find that people earned significantly fewer profit points in the malus treatment than individuals with the same degree of loss aversion in the bonus treatment. This effect declines in magnitude, but remains statistically significant when we control for baseline performance and demographics. With specification 3 as a benchmark, in the malus treatment individuals earned, on average, 1.077 points fewer per task than participants with the same ability and the same degree of loss aversion in the bonus treatment. In addition, considering the interaction effect between the malus treatment and the individual's degree of loss aversion, results show a significantly positive malus treatment effect for individuals with a low level of loss aversion, whereas the effect is significantly negative for individuals with a high degree of loss aversion. For example, specification 3 reveals that in the malus treatment individuals who rejected six lotteries in the loss aversion test accumulated 1.4 profit points per task fewer than individuals who rejected only one lottery. Our findings further show a significantly positive impact of ability, measured by the baseline performance, on profit points per task. Furthermore, specification 4 reveals that age is negatively associated with profit points.

FIGURE 3.2: Individual's degree of loss aversion and profit points



The graph plots the relationship between profit points and the degree of loss aversion with the corresponding 95% confidence intervals for both treatments. Results are based on individuals with an average baseline performance.

Figure 3.2 delivers a graphical representation of the average profit points per task in respect of the degree of loss aversion of participants in the bonus and malus treatments. The graph displays the results of specification 3. It shows that with an increasing degree of loss aversion, profits decrease in the malus condition, whereas profits are independent of loss aversion preferences in the bonus treatment. With regard to the analysis the following result is supported:

Result 1: *On average, individuals with a high level of loss aversion in the malus treatment earned significantly fewer profit points than all other individuals.*

3.5.2 Response time and deadline-dependent incentives

In both treatments, incentives were linked to a 10-second deadline. Thus, individuals received a bonus or, respectively, avoided a malus when they gave their answers within 10 seconds. Table 3.4 shows descriptive statistics of the response time and the number of received bonuses or, respectively, the number of avoided maluses for individuals with high and low degrees of loss aversion. In the malus treatment individuals with high loss averse preferences needed, on average, 9.49 seconds per task, whereas in the bonus treatment individuals with a high level of loss aversion solved the tasks, on average, in 8.06 seconds, and individuals with a low level of loss aversion entered their answers, on average, in approximately 8.4 seconds independently of the treatment. According to Mann-Whitney tests, differences in the mean response time are not significant either across treatments or across degrees of loss aversion.[7]

Table 3.4: Descriptive statistics of the response time and received bonus/avoided malus

		Malus HLA (n=13)	Malus LLA (n=17)	Bonus HLA (n=14)	Bonus LLA (n=20)
Response time	Mean	9.49	8.40	8.06	8.44
	S.D.	2.19	1.03	2.28	1.17
Received bonuses	Mean	13.31	17.65	18.00	16.80
/avoided maluses	S.D.	7.33	3.22	2.94	4.62

The table shows means and standard deviations for the response time and bonus received/malus avoided of individuals with a high level of loss aversion (HLA) and those with a low level of loss aversion (LLA).

[7] Individuals with a high degree of loss aversion in the malus treatment did not need significantly longer to respond than individuals with a low degree of loss aversion in the malus treatment (p=0.233; Mann-Whitney test) or than individuals in the bonus treatment (for individuals with a high degree of loss aversion p=0.357 and for people with a low degree of loss aversion p=0.224; Mann-Whitney test). Furthermore, there is neither a significant treatment effect between people with a low degree of loss aversion (p=0.951; Mann-Whitney test) nor a difference in the mean response time between people with different degrees of loss aversion in the bonus treatment (p=0.753; Mann-Whitney test) nor between individuals with a low degree of loss aversion in the malus treatment and individuals with a high degree of loss aversion in the bonus treatment (p=0.812; Mann-Whitney test).

Nevertheless, the differences in response times were large enough to hinder individuals with high loss averse preferences from avoiding maluses. Table 3.4 shows that participants with a low level of loss aversion could avoid, on average, 17.65 maluses, whereas individuals with a high degree of loss aversion only evaded 13.31 maluses (p=0.101; Mann-Whitney test). Furthermore, participants with a high level of loss aversion prevented fewer maluses than individuals with the same degree of loss aversion received bonuses (p=0.044; Mann-Whitney test). However, we do not find that individuals with a high degree of loss aversion avoided significantly fewer maluses than individuals with a low degree of loss aversion accumulated bonuses (p=0.129; Mann-Whitney test). Furthermore, there are no significant differences either between participants with a low level of loss aversion in the two treatments (p=0.729; Mann-Whitney test) or between individuals with different degrees of loss aversion in the bonus treatment (p=0.443; Mann-Whitney test). In addition, we do not find any significant differences between participants with a low degree of loss aversion in the malus treatment and those with a high level of loss aversion in the bonus treatment (p=0.616; Mann-Whitney test).

In Table 3.5, we examine by means of a random effects model whether the different deadline-dependent incentive schemes affected the speed of answering. The dependent variable is the response time. Independent variables and the interaction effect are the same as in the regression analysis on profit points in Section 3.5.1. Specification 1 shows that neither being in the malus treatment nor the degree of loss aversion significantly affect the response time. However, when we include the interaction effect, as in specifications 2, 3, and 4, our results show that individuals with a high degree of loss aversion in the malus treatment needed significantly more time per task than individuals with the same loss aversion level in the bonus treatment. For example, in specification 3 participants in the malus treatment were 0.785 seconds slower than individuals in the bonus treatment. Furthermore, in the malus treatment participants with a high level of loss aversion had significantly longer response times than people with a low degree of loss aversion. When we control for baseline performance and demographics, the treatment effect as well as the interaction between the malus treatment and the degree of loss aversion become marginally significant. However, the magnitude of the coefficients does not significantly change.

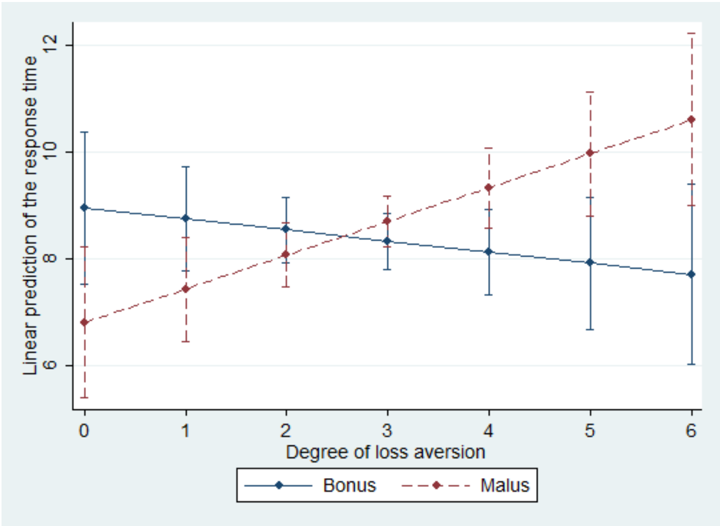
Table 3.5: Determinants of the response time

	1	2	3	4
Malus	0.586	-2.154**	-2.022*	-2.210*
	(0.421)	(1.024)	(1.119)	(1.276)
Loss aversion	0.152	-0.208	-0.210	-0.218
	(0.199)	(0.250)	(0.257)	(0.254)
Malus x loss aversion		0.842**	0.785*	0.865*
		(0.350)	(0.414)	(0.458)
Baseline performance			-0.103	-0.081
			(0.166)	(0.178)
Age				0.041
				(0.029)
Female				-0.144
				(0.470)
Years of study				-0.137
				(0.148)
Constant	7.791***	8.955***	9.647***	8.967***
	(0.601)	(0.726)	(0.872)	(1.142)
N	1278	1278	1278	1218
$R^2 overall$	0.025	0.083	0.093	0.107
Chi^2	2.484	7.639	14.366	20.383
$Sigmau$	1.682	1.598	1.593	1.655
Rho	0.541	0.516	0.514	0.530

The table shows estimates from a random effects model. Robust standard errors clustered on individual levels are in parentheses. Sample sizes differ because three students did not answer the question on how many years they had studied. As two individuals did not complete one task, the sample size differs from that in Table 3.3. Loss aversion is measured by the number of rejected lotteries in the loss aversion test. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 3.3 displays the cross-over interaction effect of the response time and the degree of loss aversion of specification 3. In the malus treatment, the response time increases with the degree of loss aversion, whereas in the bonus treatment the response time slightly decreases with higher degrees of loss aversion.

Figure 3.3: Individual's degree of loss aversion and the response time



The graph plots the relationship between the response time and the degree of loss aversion with the corresponding 95% confidence intervals for both treatments. Results are based on individuals with an average baseline performance.

Since bonus and malus payments were linked to a 10-second deadline, the main goal was to respond within this time threshold. Thus, it is worth analyzing whether the interaction between an individual's degree of loss aversion and the deadline-dependent incentive affects the probability of a participant's ability to meet the deadline and hence avoid the malus payment or receive the bonus. In Table 3.6, we present the estimates of random effects logit regressions which examine the determinants influencing the propensity of receiving a bonus or avoiding a malus. We use the same independent variables as before. The results of specification 1 show that neither being in the malus treatment nor the degree of loss aversion have a significant influence on the probability of receiving the deadline-dependent incentive. However, specification 2, which includes the interaction between the malus treatment and the degree of loss aversion, reveals that individuals with a high level of loss aversion were significantly less likely to avoid a malus than individuals with the same degree of loss aversion were likely to accumulate a bonus. When we control for the baseline performance, as in specification 3, this interaction effect decreases in magnitude but stays significant at the 5% level. Furthermore, specification 4 shows that, when we include demographic control variables like age, gender, and years of study, the coefficient of the interaction term continues to decrease in magnitude and becomes marginally significant. The odds ratio of the interaction between the malus treatment and the degree of loss aversion is 0.280 in specification 2, 0.392 in specification 3, and 0.410 in specification 4. Therefore, with specification 3 as a benchmark, for a given individual, an additional rejected lottery in the loss aversion test in the malus treatment is 0.392 times the effect of an additional rejected lottery in the bonus treatment. Additionally, without further controls, specifications 2 and 3 show a significant positive malus treatment effect for individuals with a low degree of loss aversion. Furthermore, specifications 2 and 4 reveal that the baseline performance is a significant determinant of the probability of receiving a bonus or avoiding a malus.

Table 3.6: Determinants of received bonuses/avoided maluses

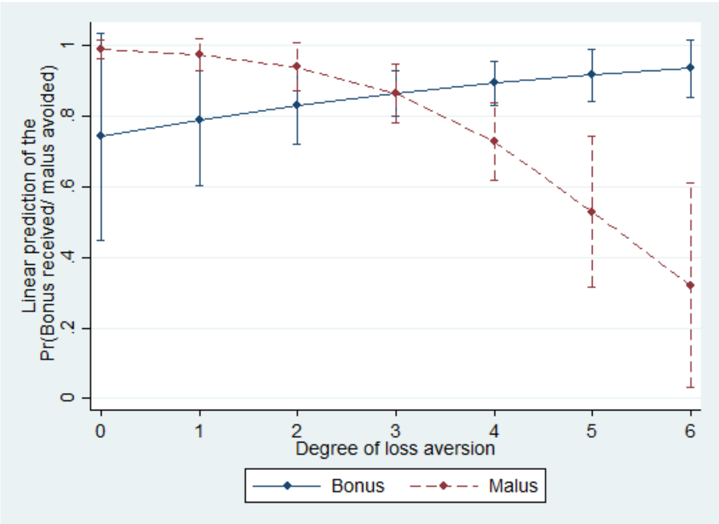
	1	2	3	4
Malus	-0.775	3.443**	2.660*	2.544
	(0.600)	(1.632)	(1.534)	(1.633)
Loss aversion	-0.292	0.299	0.257	0.190
	(0.267)	(0.310)	(0.288)	(0.285)
Malus x loss aversion		-1.275***	-0.936**	-0.891*
		(0.464)	(0.473)	(0.498)
Baseline performance			0.428***	0.405***
			(0.112)	(0.117)
Age				-0.057
				(0.038)
Female				0.221
				(0.526)
Years of study				0.002
				(0.197)
Constant	3.968***	1.973*	-0.849	0.646
	(1.051)	(1.111)	(1.086)	(1.720)
lnsig2u	1.514***	1.301***	0.842***	0.800***
	(0.255)	(0.225)	(0.283)	(0.300)
N	1278	1278	1278	1218
Individuals	64	64	64	61
Chi^2	3.340	12.307	32.476	35.358
Sigmau	2.132	1.917	1.524	1.492
Rho	0.580	0.528	0.414	0.403

The table shows estimates from a random effects logit model. Robust standard errors clustered on individual levels are in parentheses. Sample sizes differ because three students did not answer the question on how many years they had studied. As two individuals did not complete one task, the sample size differs from the one in Table 3.3. Loss aversion is measured by the number of rejected lotteries in the loss aversion test. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The results of specification 3 are also summarized in Figure 3.4. The graph displays the predicted probabilities and gives a visual representation of the interaction effect between the treatments and the degree of loss aversion. It shows that the probability of avoiding a malus declines with the extent of the degree of loss aversion, whereas the probability of receiving a bonus is independent of an individual's degree of loss aversion. Our second result can be summarized as follows:

Result 2: *On average, individuals with a high level of loss aversion in the malus treatment needed more time to respond and thus they were more likely to exceed the 10-second deadline and suffered more malus payments than all other individuals.*

FIGURE 3.4: Individual's degree of loss aversion and the probability of receiving a bonus/avoiding a malus



The graph plots the relationship between the probability of receiving a bonus/avoiding a malus and the degree of loss aversion with the corresponding 95% confidence intervals for both treatments. Results are based on individuals with an average baseline performance.

3.5.3 Correct answers

Table 3.7 shows that, on average, participants with a high level of loss aversion in the malus treatment reported 15.46 correct answers, whereas all other individuals completed approximately 18 tables correctly. Applying Mann-Whitney tests, we find that in the malus treatment participants with a high level of loss aversion solved significantly fewer answers correctly than individuals with a low degree of loss aversion (p=0.009) and individuals in the bonus treatment (for participants with a high degree of loss aversion p=0.040 and a low degree of loss aversion p=0.034). We do not find any significant results when comparing the mean number of correct answers of participants with a low degree of loss aversion in the bonus treatment with that of those with a low degree of loss aversion in the malus treatment (p=0.696; Mann-Whitney test). Neither do we find any significant differences between the mean number of correct answers of individuals with high and those with low loss aversion preferences in the bonus treatment (p=0.943; Mann-Whitney test). Furthermore, individuals with a low level of loss aversion in the malus treatment did not give significantly more answers correctly than participants with high levels of loss aversion in the bonus treatment (p=0.596; Mann-Whitney test).

Table 3.7: Descriptive statistics of correct answers

	Malus HLA (n=13)	Malus LLA (n=17)	Bonus HLA (n=14)	Bonus LLA (n=20)
Mean	15.46	18.18	17.50	17.75
S.D.	3.45	2.21	3.55	2.49

The table shows means and standard deviations for correct answers of individuals with a high level of loss aversion (HLA) and those with a low level of loss aversion (LLA).

The question of whether there is a treatment effect on correctly solved tasks in terms of individuals' degree of loss aversion is also addressed in Table 3.8. It reports the estimates of a random effects logit model. Specification 1 shows that individuals with a higher degree of loss aversion were less likely to give a correct answer than those with a low degree of loss aversion. In specification 2, when we include the interaction term between the malus treatment and the degree of loss aversion, our results suggest that compared with the bonus treatment, an additional rejected lottery in the loss aversion test decrease the propensity of entering a correct answer in the malus treatment. In specification 3, where we also control for the baseline performance, the coefficient on the interaction between the degree of loss aversion and the malus treatment declines in size but stays statistically significant. As expected, the baseline performance is positively related to the propensity of giving a correct answer. Specifications 2 and 3 further suggest that compared with the bonus treatment, individuals with a low level of loss aversion in the malus treatment solved more tasks correctly. However, when we add demographic control variables like age, gender, and years of study, the coefficient of the interaction effect declines and becomes insignificant. The odds ratio for the interaction term is 0.564 in specification 2, 0.657 in specification 3, and 0.717 in specification 4. This tells us that, for example, in specification 3 the effect of an increase in loss aversion for individuals in the malus treatment is 0.657 times the effect of individuals with the same degree of loss aversion in the bonus treatment.

TABLE 3.8: Determinants of correct answers

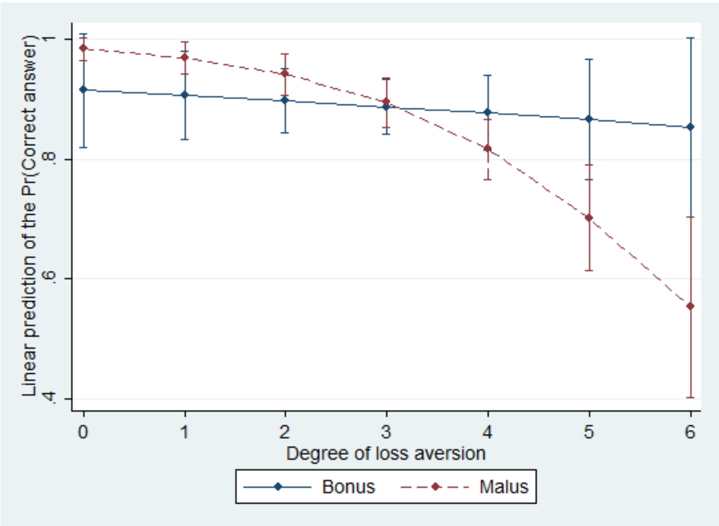
	1	2	3	4
Malus	-0.292	1.672**	1.318*	1.001
	(0.323)	(0.814)	(0.719)	(0.806)
Loss aversion	-0.345**	-0.092	-0.093	-0.104
	(0.137)	(0.196)	(0.137)	(0.143)
Malus x loss aversion		-0.573**	-0.420**	-0.332
		(0.228)	(0.206)	(0.227)
Baseline performance			0.253***	0.254***
			(0.064)	(0.061)
Age				-0.023
				(0.023)
Female				-0.341
				(0.293)
Years of study				-0.044
				(0.073)
Constant	3.556***	2.682***	0.978*	1.801*
	(0.517)	(0.684)	(0.572)	(0.961)
lnsig2u	0.028	-0.139	-0.770**	-0.799**
	(0.312)	(0.381)	(0.364)	(0.367)
N	1278	1278	1278	1218
Individuals	64	64	64	61
Chi^2	8.690	14.910	39.964	41.671
Sigmau	1.014	0.933	0.680	0.671
Rho	0.238	0.209	0.123	0.120

The table shows estimates from a random effects logit model. Robust standard errors clustered on individual levels are in parentheses. Sample sizes differ because three students did not answer the question on how many years they had studied. As two individuals did not complete one task, the sample size differs from the one in Table 3.3. Loss aversion is measured by the number of rejected lotteries in the loss aversion test. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In addition, the relationship between the probability of solving a task correctly and the degree of loss aversion is also shown in Figure 3.5. It indicates that individuals with a higher degree of loss aversion in the malus treatment were less likely to enter a correct answer, whereas the probability of giving a correct answer in the bonus treatment was independent of the individual's degree of loss aversion. According to our analysis of correct answers we can report the following result:

Result 3: *Individuals with a high level of loss aversion in the malus treatment solved fewer tasks correctly than all other individuals. However, when we control for individual baseline performance and other demographics, the interaction effect between the malus treatment and the degree of loss aversion becomes insignificant.*

FIGURE 3.5: Individual's degree of loss aversion and the probability of giving a correct answer



The graph plots the relationship between the probability of giving a correct answer and the degree of loss aversion with the corresponding 95% confidence intervals for both treatments. Results are based on individuals with an average baseline performance.

3.5.4 Results across stages

We also compare the average response time and the share of correct answers across stages. In the first and third stages participants had to count the zeros in five tables. In these stages, no deadline-dependent incentive scheme was applied. The first stage was introduced to measure baseline performance and the last stage for identifying possible fatigue effects and reactions to the removal of the deadline-dependent incentive scheme. Results show that, on average, individuals' response time was faster in the main stage, independently of their degree of loss aversion and the treatment (main stage compared with stage 1 $p<0.001$, main stage compared with stage 3 $p<0.001$; Wilcoxon signed rank test). When we compare the first to the third stage, the data suggests a significant decrease in the response time in both treatments ($p<0.001$ for both treatments; Wilcoxon signed rank test). Hence, we do not find any fatigue effects but rather learning effects or a lasting effect of the deadline-dependent incentive scheme from the previous stage. We do not find any significant results when comparing the different average response times between stages of participants with a high and a low degree of loss aversion.

When we compare the share of correct answers across stages, two-sided Wilcoxon signed rank tests show that independently of the treatment individuals with a low degree of loss aversion have a significant higher share of correct answers in the third stage than in the first (p=0.018 for the bonus treatment and p=0.003 for the malus treatment) or in the main stage (p=0.048 for the bonus treatment and p=0.033 for the malus treatment). In addition, participants displaying high loss averse preferences in the malus treatment have a significant higher share of correct answers in the third stage than in the main stage (p=0.080; Wilcoxon signed rank test). However, when comparing the share of correct answers in the first stage with those in the main stage of individuals with a high degree of loss aversion in the malus treatment we do not find any significant results. Altogether, deadline-dependent incentive schemes lead to faster response times than environments without time pressure. In this respect, our experiment complements that of Kocher and Sutter (2006). Further, our results suggest learning effects in respect of both quality and quantity.

3.6 Discussion and conclusion

Many economically relevant activities such as financial decisions, finishing a construction project or working for piece rate wages are executed under notable time pressure. In these situations, incentivized deadlines are often the reason why people feel time-pressured. Bonus payments are commonly offered for finishing a project within a specific time-frame and malus payments are imposed for not meeting a deadline. Given the importance of examining the efficiency of contracts in a time-pressured environment, this paper addresses the question of whether individuals' loss averse preferences influence performance under deadline-dependent bonus and malus incentive schemes.

Several studies show that effort provision is higher under malus contracts than under bonus contracts (e.g., Hannan et al. 2005, Church et al. 2008, Brooks et al. 2012). However, bonus incentives are still the most common variable compensation form. By conditioning incentives on time and considering individuals' degrees of loss aversion, this study adds to the discussion on the predominance of bonus incentive schemes. In our experiment we linked incentives to a deadline and examine the distinct effects of bonus and malus contracts on performance under time pressure in a real effort task. In general, individuals who are more loss averse have a higher need to avoid malus payments. Thus, they are more likely to suffer from possible distractions and to choke under pressure. Indeed, our findings provide evidence that an individual's degree of loss aversion is an indicator for performance differences under deadline-dependent incentives. Our results show that individuals with a high level of loss aversion performed worse under a malus contract than all other individuals. These findings are driven by both speed and quality and are thus in line with our hypotheses. We suggest that the choking effect in the time dimension of highly loss averse individuals in the malus treatment is caused by the fact that the malus scheme is linked to a deadline. Furthermore, we argue that the decrease in quality is explainable by the predictions of distraction theory which postulates that the pressure variable shifts the attention away from the actual tasks to worrisome thoughts. In our case, the deadline-dependent malus scheme and in particular the countdown at the computer screen might have created a distracting environment, which drew the attention from the zero-counting task to the deadline. Additionally, we observe that individuals with a very low degree of loss aversion increased their performance in the malus treatment. This result supports previous findings that malus contracts are more effective (Hannan et al. 2005, Church et al. 2008, Brooks et al. 2012). Furthermore, in contrast to the malus contract, performance under a bonus contract was independent of the individual's degree of loss aversion.

The above results are in line with the findings of Payne et al. (1996), who examined the detrimental effects of time-dependent incentives on decision quality. However, Kocher and Sutter (2006) showed that time-dependent incentives lead to faster decisions without a quality decrease. Payne et al. (1996) as well as Kocher and Sutter (2006) implemented a time-dependent factor in the payoff which implied that faster decisions were rewarded, whereas slower decisions were punished. Compared with these studies, we investigated the impact of bonus and malus contracts separately. This allowed us to analyze which contract form is decisive for performance differences and whether an individual's degree of loss aversion is a determining factor for performance differences under the two incentive schemes. In addition, since companies often use deadlines to achieve short processing times, we linked our incentives to a strict deadline instead of applying a time-dependent factor to the payoff. Another contribution of our study to the literature concerns the task participants had to fulfill. Whereas Payne et al. (1996) faced their participants with risky decisions and Kocher and Sutter (2006) with a strategic environment, we used a riskless and non-strategic real effort task.

Recent research reveals hidden costs of malus contracts. For example, cheating is more prominent under a loss frame (Grolleau et al. 2014). In addition, Brooks et al. (2013) declare malus contracts as a risky contract choice, because the effectiveness might rely on the chosen threshold around which earnings are framed. If the threshold is very high, malus contracts can have counterproductive effects. Our research adds to this by showing that, under time pressure, malus contracts work worse for individuals with a high level of loss aversion. Even though these factors need to be considered for contract design, there are various applications where malus incentives seem to be a fitting mechanism. For example, malus points for wrong answers can prevent students from guessing in multiple choice exams and malus components in insurance policies can motivate people to adopt more prudent conduct. Likewise, banks have started including malus incentives in their managers' remuneration systems to prevent excessively risky behavior. However, in none of these examples is the malus incentive linked to a deadline which may be the reason for the positive effect.

Although the analysis improves our understanding of performance differences under time pressure, it raises a number of new questions that should be explored by future research. In our experiment incentives were directly linked to a deadline, and the evidence presented does not allow us to make a final assessment of the sole effect of time pressure. Kocher and Sutter (2006) analyzed the influence of pure time pressure on the quality of strategic decision making and found that time pressure leads to worse decisions when there is no time-dependent incentive scheme in place. Similarly to Kocher and Sutter (2006), a treatment with a non-incentivized but nevertheless challenging deadline would be appropriate for isolating the pure time pressure effect. Future research should address this issue in a non-strategic setting such as ours. Prospectively, the incentive scheme itself should be considered in more detail, whether it is framed or deadline-dependent or a combination of different elements. As observed in other studies (Deffenbacher 1978, Baumeister et al. 1993), the interaction between individual characteristics and situational elements can be a powerful determinant of performance. In relation to the topic of choking under pressure, research should address the influence of other pressure variables such as social comparison and the threat of a stereotype, in combination with different incentive schemes. Furthermore, to gain more insights into people's mental and physical functions in time-pressured environments, research should try to measure people's stress levels while working under different incentives with the help of reliable physical measurements such as blood pressure level and pulse.

3.7 References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–492.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *Review of Economic Studies*, 76(2):451–469.

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3):610–620.

Baumeister, R. F., Heatherton, T. F., and Tice, D. M. (1993). When ego threats lead to self-regulation failure: Negative consequences of high self-esteem. *Journal of Personality and Social Psychology*, 64(1):141–156.

Baumeister, R. F. and Showers, C. J. (1986). A review of paradoxical performance effects: Choking under pressure in sports and mental tests. *European Journal of Social Psychology*, 16(4):361–383.

Beilock, S. L. and Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General*, 130(4):701–725.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5):841–877.

Bracha, A. and Fershtman, C. (2013). Competitive incentives: Working harder or working smarter? *Management Science*, 59(4):771–781.

Brink, A. G. and Rankin, F. W. (2013). The effects of risk preference and loss aversion on individual behavior under bonus, penalty, and combined contract frames. *Behavioral Research in Accounting*, 25(2):145–170.

Brooks, R. R. W., Stremitzer, A., and Tontrup, S. (2012). Framing contracts: Why loss framing increases effort. *Journal of Institutional and Theoretical Economics*, 168(1):62–82.

Brooks, R. R. W., Stremitzer, A., and Tontrup, S. (2013). Stretch it but don't break it: The hidden risk of contract framing. *UCLA School of Law, Law-Econ Research Paper No. 13-22*.

Cavanaugh, M. A., Boswell, W. R., Roehling, M. V., and Boudreau, J. W. (2000). An empirical examination of self-reported work stress among U.S. managers. *Journal of Applied Psychology*, 85(1):65–74.

Church, B. K., Libby, T., and Zhang, P. (2008). Contracting frame and individual behavior: Experimental evidence. *Journal of Management Accounting Research*, 20(1):153–168.

Cubitt, R. P., Starmer, C., and Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, 1(2):115–131.

DeCaro, M. S., Thomas, R. D., Albert, N. B., and Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. *Journal of Experimental Psychology: General*, 140(3):390–406.

Deffenbacher, J. L. (1978). Worry, emotionality, and task-generated interference in test anxiety: An empirical test of attentional theory. *Journal of Educational Psychology*, 70(2):248–254.

Dohmen, T. J. (2008). Do professionals choke under pressure? *Journal of Economic Behavior & Organization*, 65(3-4):636–653.

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

Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317.

Fehr, E., Goette, L., and Lienhard, M. (2008). Loss aversion and effort: Evidence from a field experiment. *Working Paper, University of Zurich.*

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

Fryer, R. G., Levitt, S. D., List, J., and Sadoff, S. (2013). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *NBER Working Paper 18237.*

Gächter, S., Johnson, E. J., and Herrmann, A. (2010). Individual-level loss aversion in riskless and risky choices. *CeDEx Discussion Paper Series No. 2010-20.*

Greiner, B. (2004). An online recruitment system for economic experiments. In Kremer, K. and Macho, V., editors, *Forschung und Wissenschaftliches Rechnen*, volume 63, pages 79–93. Gesellschaft für wissenschaftliche Datenverarbeitung, Göttingen.

Grolleau, G., Kocher, M. G., and Sutan, A. (2014). Cheating and loss aversion: Do people lie more to avoid a loss? *CESifo Working Paper Series No. 4965.*

Hannan, L. R., Hoffman, V. B., and Moser, D. V. (2005). Bonus versus penalty: Does contract frame affect employee effort? In Zwick, R. and Rapoport, A., editors, *Experimental Business Research*, volume 63, pages 151–169. Springer, Dordrecht, The Netherlands.

Hilken, K., Rosenkranz, S., De Jaegher, K., and Jegers, M. (2013). Reference points, performance and ability: A real effort experiment on framed incentive schemes. *Tjalling C. Koopmans Research Institute Discussion Paper Series No. 13-15.*

Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.

Imas, A., Sadoff, S., and Samek, A. S. (2015). Do people anticipate loss aversion? *Working Paper SSRN 2411023.*

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–292.

Kocher, M. G. and Sutter, M. (2006). Time is money - Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization*, 61(3):375–392.

Luft, J. (1994). Bonus and penalty incentives contract choice by employees. *Journal of Accounting and Economics*, 18(2):181–206.

Payne, J. W., Bettman, J. R., and Luce, M. F. (1996). When time is money: Decision behavior under opportunity-cost time pressure. *Organizational Behavior and Human Decision Processes*, 66(2):131–152.

Quidt, J. (2014). Your loss is my gain: A recruitment experiment with framed incentives. *Working Paper SSRN 2418218.*

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, 68(5):1281–1292.

Sanders, S. and Walia, B. (2012). Shirking and "choking" under incentive-based pressure: A behavioral economic theory of performance production. *Economics Letters*, 116(3):363–366.

Shurchkov, O. (2012). Under pressure: Gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10(5):1189–1213.

Smith, J. (2013). Peers, pressure, and performance at the national spelling bee. *Journal of Human Resources*, 48(2):265–285.

3.8 Appendix A: Experimental instructions for the malus treatment

(Original instructions are in German)

General instructions

Welcome to this scientific experiment. Please read the instructions carefully. At the end of the experiment you will be privately paid in cash according to your decisions. For your punctual arrival, you receive compensation of 6 CHF. When you take the tasks seriously, you can achieve corresponding earnings in Swiss francs. Your answers will be treated in strict confidence, i.e. all data are evaluated anonymously by a third person, who was not present in the laboratory.

Please remember that you are not allowed to speak during the experiment. The use of mobile phones, smartphones, tablet-PCs, etc. is forbidden. Any interferences will mean your exclusion from the experiment and the loss of all earnings.

If you have any questions during the experiment, please raise your hand. The experimenter will come to you and answer your questions.

The experiment consists of four parts and a questionnaire. For parts 1 and 2 you will receive written instructions. The instructions for parts 3 and 4 will be displayed directly on the screen. All four parts will be paid out in cash after the experiment.

During parts 1, 2, and 3 of the experiment, we will talk about points instead of Swiss francs. Your payment is first calculated in points. At the end it will be converted into Swiss francs, and the exchange rate is as follows:

$$10 \text{ points} = 0.4 \text{ CHF}.$$

At the end of the experiment we would kindly ask you to fill in the questionnaire. The answers given in the questionnaire have no influence on your cash earnings in the experiment. Again, all information is treated in strict confidence.

Please read the instructions for part 1 of the experiment and answer the comprehension questions on the screen.

Thank you for your participation. We hope you enjoy the experiment.

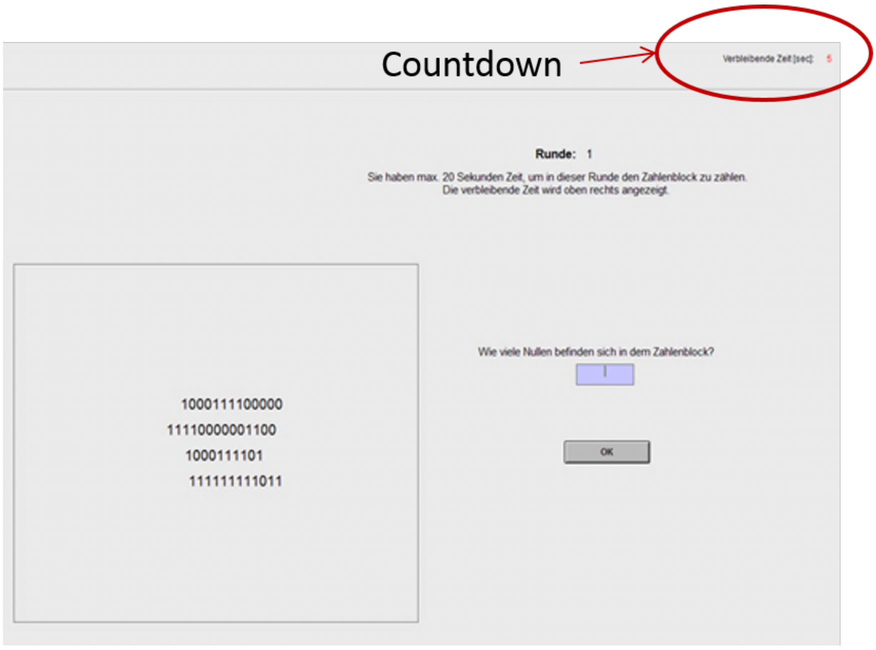
Guidance on part 1

In the first part of the experiment you are asked to work on five tasks. For each task you have a maximal time of 20 seconds. One task consists of a table, out of which you have to determine the correct number of the digit "0". Once you have determined the number of the digit "0" in the table, you should enter the number in the corresponding field and confirm your entry by clicking on the OK-button.

For every **correct answer you receive 15 points**, for every **wrong answer you receive 6 points** and for every unsolved task your receive 0 points. Once you have confirmed your entry, you will receive feedback on whether you have solved the task correctly or not. Thereafter the next task starts with a new table. Your remaining time in every single task is displayed in red at the top right-hand of your screen.

At the end of the first part of the experiment you will receive feedback on the number of correctly solved tasks, the number of incorrectly solved tasks, and the number of the unsolved tasks as well as the resulting payment in points.

Here you can see a screen shot of the task:



After part 1 you will receive instructions for the next part of the experiment.

Instructions on part 2

In the second part of the experiment you are asked to work on a total of 20 tasks. You have a maximum time of 20 seconds for each task. Again, one task consists of a table, out of which you have to determine the right number of the digit "0". The input screen equals the one of the first part of the experiment. After you have counted the right number of the digit "0" in a square, please enter it into the corresponding field. Then confirm it by clicking on the OK-button. The time remaining to solve the task is displayed in red at the top right-hand corner of the computer screen.

Your payments

The general rules are as follows:

- For each correct answer you receive 15 points.
- For each wrong answer you receive 6 points.
- If you do not enter anything within the 20 seconds, you get 0 points.
- If you need **more than 10 seconds** to solve a task, you receive **a malus of 5 points**.

Once you have confirmed your entry you will receive feedback. The feedback tells you whether you have solved the task correctly and whether you have received a malus of 5 points. Then the next task starts on your screen.

Examples:

- You solve a task correctly but you need more than 10 seconds. In this case, you receive 15 points for your correct answer and a malus of 5 points for needing more than 10 seconds to solve the task.
- You solve a task incorrectly and need more than 10 seconds. In this case, you receive 6 points for your wrong answer and a malus of 5 points for needing more than 10 seconds to solve the task.

After finishing the last task, you will receive feedback on the number of correctly solved tasks, the number of incorrectly solved tasks, the number of unsolved tasks and the number of the received malus points as well as your accumulated points for this part of the experiment.

Now please solve the comprehension questions on the screen and then start working on the tasks.

After part 2, the instructions for part 3 will appear directly on your computer screen. The instructions for part 4 will also be displayed on the screen. After part 4 you can see your earnings for each part of the experiment as well as your accumulated earnings in Swiss francs on the screen.

Part 3 was equal to part 1

Part 4

Below you see 6 different lottery decision situations, where you can either choose to accept or reject the lottery. You have to decide for all 6 situations whether you accept or reject the corresponding lottery. At the end of the experiment one lottery will be randomly determined and paid out.

1. Win 6 CHF with a probability of 50% & lose 2 CHF with a probability of 50%.
   Do you participate in this lottery?
   ☐ Accept
   ☐ Reject

2. Win 6 CHF with a probability of 50% & lose 3 CHF with a probability of 50%.
   Do you participate in this lottery?
   ☐ Accept
   ☐ Reject

3. Win 6 CHF with a probability of 50% & lose 4 CHF with a probability of 50%.
   Do you participate in this lottery?
   ☐ Accept
   ☐ Reject

4. Win 6 CHF with a probability of 50% & lose 5 CHF with a probability of 50%.
   Do you participate in this lottery?
   ☐ Accept
   ☐ Reject

5. Win 6 CHF with a probability of 50% & lose 6 CHF with a probability of 50%.
   Do you participate in this lottery?
   ☐ Accept
   ☐ Reject

6. Win 6 CHF with a probability of 50% & lose 7 CHF with a probability of 50%.
   Do you participate in this lottery?
   ☐ Accept
   ☐ Reject

Once you have made your decisions please confirm them with the OK button. After you have clicked the button, you cannot change your decision.

Thank you for participating in this experiment!

3.9 Appendix B: Further results

In the subsequent sections, we analyze profit points, the response time, received bonuses or avoided maluses and correct answers for all 20 tasks. For both the malus and the bonus treatment, we separately present the results for individuals with a high level of loss aversion and those with a low level of loss aversion.

3.9.1 Profit points over time

Figure 3.6 suggests that individuals with a high degree of loss aversion in the malus treatment accumulated fewer profit points over the 20 periods than other individuals. Comparing profit points of individuals with high loss averse preferences in both treatments, we find that in six tasks people in the malus treatment earned significantly fewer profit points than those in the bonus treatment.[8] Analyzing profit points in the malus treatment, we find that participants with a high level of loss aversion earned in eight tasks significantly fewer profit points than people with a low level of loss aversion.[9] Furthermore, participants with a high level of loss aversion in the malus treatment collected significantly fewer profit points in six tasks than individuals with a low level of loss aversion in the bonus treatment.[10] Only in two tasks did participants with a low level of loss aversion earn significantly less in the bonus than in the malus treatment.[11] Except for one task, we do not find any significant differences between individuals with a high degree and those with a low degree of loss aversion in the bonus treatment.[12] Furthermore, we do not find any significant differences between individuals with a high degree of loss aversion in the bonus treatment and those with a low degree of loss aversion in the malus treatment.

[8] For tasks 3 (p=0.034), 9 (p=0.045), 10 (p=0.085), 15 (p=0.019), 18 (p=0.020), 20 (p=0.096); Mann-Whitney test.

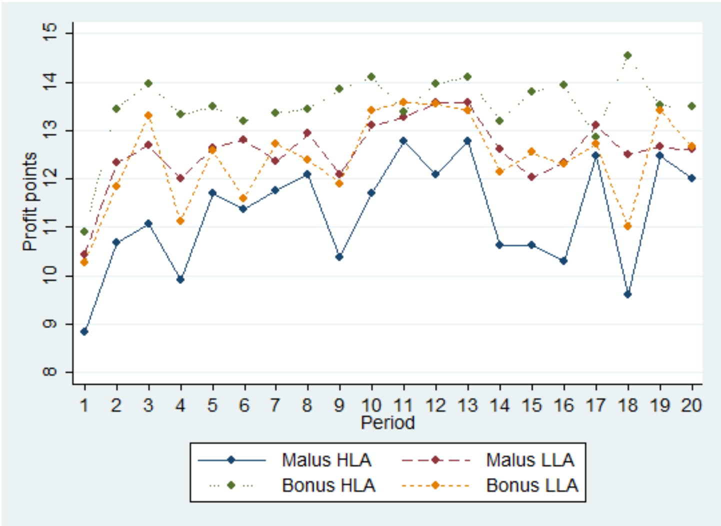
[9] For tasks 2 (p=0.060), 3 (p=0.024), 4 (p=0.046), 6 (p=0.095), 10 (p=0.043), 12 (p=0.026), 14 (p=0.025), and 18 (p=0.008); Mann-Whitney test.

[10] For tasks 3 (p=0.001), 10 (p=0.007), 12 (p=0.044), 14 (p=0.071), 15 (p=0.08,) and 16 (p=0.062); Mann-Whitney test.

[11] For tasks 6 (p=0.067), and 18 (p=0.016); Mann-Whitney test.

[12] For task 18 (p=0.044); Mann-Whitney test.

FIGURE 3.6: Profit points earned over 20 periods



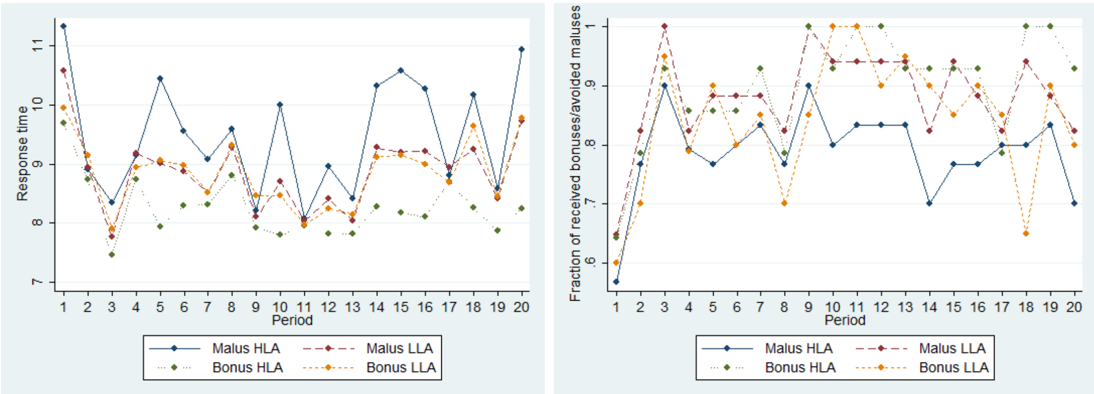
3.9.2 Response time and deadline-dependent incentives over time

Graph (a) in Figure 3.7 shows that participants with a high level of loss aversion in the malus treatment needed more time for solving the tasks than other individuals. Individuals with a high degree of loss aversion needed in two tasks significantly more time in the malus treatment than in the bonus treatment.[13] Within the malus treatment, individuals with a low level of loss aversion reported their answers significantly faster in two tasks than individuals with a high level of loss aversion.[14] The differences between individuals with a high level of loss aversion in the malus treatment and individuals with a low degree of loss aversion in the bonus treatment are significant for five tasks.[15] Within the bonus treatment, there are only significant differences between participants with high and those with low degrees of loss aversion in three tasks.[16] There are no significant treatment effects between participants with a low level of loss aversion. Neither do we find any significant differences between individuals with a high degree of loss aversion in the bonus treatment and individuals with a low degree of loss aversion in the malus treatment. Graph (b) in Figure 3.7 displays how many bonuses participants received or how many maluses they avoided. It indicates that individuals with a high degree of loss aversion indeed suffered more malus payments than other individuals. Treatment differences

[13]For tasks 5 (p=0.081) and 20 (p=0.037); Mann-Whitney test.
[14]For tasks 5 (p=0.023), and 10 (p=0.042); Mann-Whitney test.
[15]For tasks 1 (p=0.039), 5 (p=0.060), 10 (p=0.015), 14 (p=0.083), and 15 (p=0.083); Mann-Whitney test.
[16]For tasks 10 (p=0.080), 18 (p=0.069), and 20 (p=0.086); Mann-Whitney test.

between participants with a high level of loss aversion are significant for 10 tasks.[17] In 12 tasks participants who have a low level of loss aversion could evade significantly more maluses than individuals with a high level of loss aversion.[18] Furthermore, there are significant treatment differences between high loss averse individuals in the malus treatment and individuals with a low degree of loss aversion in the bonus treatment in seven tasks.[19] Apart from task 18, we do not find any significant difference between the number of received bonuses and the number of avoided maluses of individuals with a low degree of loss aversion.[20] Task 18 is also the only significant case, when we compare the received bonuses of individuals with a high level of loss aversion with those with a low level of loss aversion.[21] We do not find any significant results when we compare participants with a low degree of loss aversion in the malus treatment with those with a high level of loss aversion in the bonus treatment.

FIGURE 3.7: Response time and bonus received/malus avoided over 20 periods



(a) Graph

(b) Graph

3.9.3 Correct answers over time

Figure 3.8 indicates that participants with a high level of loss aversion in the malus treatment gave fewer correct answers than other individuals. However, the Mann-Whitney tests show that there are no significant differences in most tasks. Only in three tasks, did participants with a high level of loss aversion enter significantly more incorrect answers

[17] For tasks 9 (p=0.061), 10 (p=0.055), 11 (p=0.027), 12 (p=0.027), 14 (p=0.023), 15 (p=0.023), 16 (p=0.055), 18 (p=0.012), 19 (p=0.061), and 20 (p=0.023); Mann-Whitney test.

[18] For tasks 3 (p=0.040), 5 (p=0.092), 9 (p=0.040), 10 (p=0.030), 11 (p=0.075), 12 (p=0.075), 13 (p=0.075), 14 (p=0.097), 15 (p=0.011), 16 (p=0.092), 18 (p=0.030), and 20 (p=0.097); Mann-Whitney test.

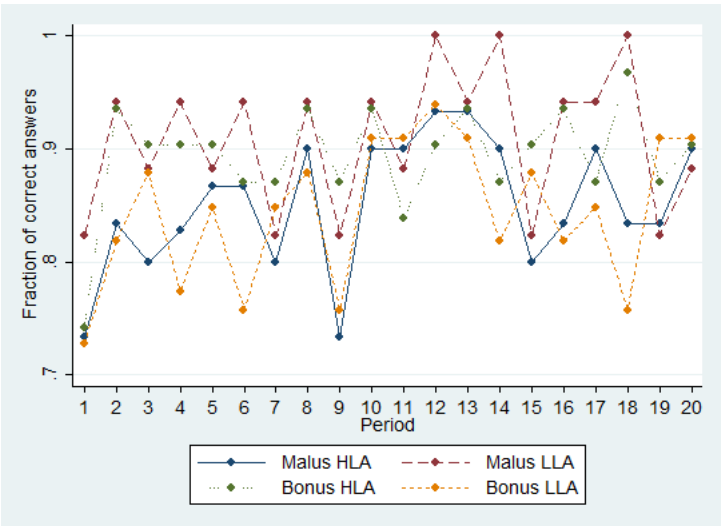
[19] For tasks 5 (p=0.054), 10 (p=0.003), 11 (p=0.009), 13 (p=0.047), 14 (p=0.020), 15 (p=0.053), and 16 (p=0.054); Mann-Whitney test.

[20] For task 18 (p=0.034); Mann-Whitney test.

[21] For task 18 (p=0.014); Mann-Whitney test.

in the malus treatment than in the bonus treatment.[22] Within the malus treatment, individuals with lower levels of loss aversion preferences gave significantly more correct answers in six tasks than individuals with a high degree of loss aversion.[23] In addition, participants with a low degree of loss aversion in the bonus treatment gave more accurate answers in two tasks than participants with a high degree of loss aversion in the malus treatment.[24] Also in two tasks, participants with a high level of loss aversion in the bonus treatment entered significantly fewer correct answers than individuals with a low level of loss aversion in the malus treatment.[25] Furthermore, we only find significant differences between individuals with a high and those with a low degree of loss aversion in one task in the bonus treatment.[26] However, we do not find any significant differences between individuals with a low degree of loss aversion in both treatments.

FIGURE 3.8: Correct answers over 20 periods



[22]For tasks 9 (p=0.055), 15 (p=0.061), and 18 (p=0.055); Mann-Whitney test.
[23]For tasks 2 (p=0.075), 4 (p=0.058), 12 (p=0.100), 14 (p=0.040), 16 (p=0.075), and 18 (p=0.006); Mann-Whitney test.
[24]For tasks 3 (p=0.009) and 12 (p=0.075); Mann-Whitney test.
[25]For tasks 12 (p=0.048) and 14 (p=0.020); Mann-Whitney test.
[26]For task 12 (p=0.033); Mann-Whitney test.

4 Essay 3: Gender and performance spillover effects of symbolic recognition at school

Andrea Essl *

Abstract

This paper examines the questions of whether symbolic recognition in one task influences the performance in a subsequent unrelated task, and if so, how the effect differs with positive, negative or no symbolic recognition. Furthermore, I analyze how the taste for recognition varies by gender. In an artefactual field experiment, secondary school students had to work on two different tasks. In the first task, they had to guess the number of peas in a bowl and in the second task, they had to cut out flyers. In the experimental treatment, the students received unannounced performance feedback after the estimation task. The top third was rewarded with a smiley-sticker, the bottom third received a frowny-sticker, and the intermediate third did not receive any symbolic recognition. The students in the control treatment received no symbolic recognition after the estimation task. I find that the response to different symbolic recognition types is heterogeneous across genders. In the flyer-cutting task following the estimation task, the female non-recipients as well as the females who received a smiley-sticker in the estimation task significantly outperformed females in the control group. I do not find spillover effects of the different recognition types on males' performance.

Keywords: symbolic recognition, performance spillover effects, gender differences, experimental economics

* Andrea Essl (andrea.essl@iop.unibe.ch) is from the Institute of Organization and Human Resource Management, University of Bern. I am grateful to Roja Abbassi who conducted the experiment as part of her master thesis and the teachers and students of the school Möriken-Wildegg. Lucian Hunger, Katja Rubin and Flavio von Rickenbach provided excellent research assistance. I would like to thank Frauke von Bieberstein, Stefanie Jaussi, Claudia Vogel and participants at NIBS (University of Nottingham) for helpful comments. All errors are mine.

4.1 Introduction

"People may take a job for more money, but they often leave it for more recognition."
(Bob Nelson)

According to the business book *1001 Ways to Reward Employees* by Nelson (2006), it is not a bonus that motivates an employee, but it is recognition. To increase employees' motivation, many companies provide non-monetary incentives like *Employee of the Month* awards. Thinslices and McDonald's are two examples of firms applying the *Employee of the Month* award. Symbolic expressions of recognition are not only common in the business world, but also a popular part of school traditions. Schools make use of symbolic recognition, such as gold stars, certificates, and prizes, for good performance to foster confidence and promote good study habits.

While standard economic models postulate that rewarding individuals with monetary incentives is an effective means of motivating people (see Prendergast 1999), recent research has shown that non-monetary recognition programs are important and cost-efficient alternatives or complements (e.g., Kosfeld and Neckermann 2011, Bradler et al. 2013, Ashraf et al. 2014, Neckermann et al. 2014). Whereas the existing literature provides fairly consistent results that ex ante mentioned recognition programs increase individuals' performance, much less is known about the spillover effect of different recognition types on the performance in a subsequent unrelated task. In this study, I investigate this question and additionally focus on how the response differs if participants receive positive, negative or no symbolic recognition. To analyze these issues, I conducted an experiment with 138 secondary school students. Pupils are a suitable sample for examining the spillover effect of symbolic recognition on performance for two reasons. First, students take several subjects that are unrelated to each other (e.g., math and languages). Therefore, they know how to deal with a multitask environment. Second, recognition programs are common tools for rewarding academic achievements; thus, school students are used to symbolic recognition. In addition, several studies have indicated that recognition programs and relative performance feedback have a positive effect on student's test scores (e.g., Azmat and Iriberri 2010, Tran and Zeckhauser 2012, Levitt et al. 2012). In the experiment of this study, the students were required to work on two different tasks. In the first task, they had to estimate the number of peas in a bowl and in the second task, they had to cut out advertising flyers for a university orchestra concert. In the experimental treatment, the participants received unannounced performance feedback after the estimation task. The top third was rewarded with a smiley-sticker, the bottom third received a frowny-sticker, and the intermediate third did not receive symbolic recognition, but the information that they had performed averagely. The students in the control treatment received no performance feedback after the estimation task.

A growing body of research has suggested that gender differences determine different reactions to relative performance feedback. Although several studies have revealed that men perform significantly better than women in tournaments (Gneezy et al. 2003, Gneezy and Rustichini 2004), other papers have suggested that gender differences in competitive environments are sensitive to the task (e.g., Günther et al. 2010, Shurchkov 2012). These studies have shown that men outperform women in stereotypically male tasks, whereas women perform equally or better than men in stereotypically female tasks. I extend the discussion by examining whether there is a gender-specific spillover effect of different symbolic recognition types between two unrelated tasks. Neither relative symbolic performance feedback nor its spillover effect has previously been explored with a focus on gender differences.

I am only aware of two studies dealing with the spillover effects of awards (Neckermann et al. 2014, Bradler et al. 2013). Both papers showed that unannounced positive recognition significantly improved subsequent performance and that this performance increase was mainly driven by individuals who were not recognized. My contribution expands on Bradler et al. (2013) and Neckermann et al. (2014) by investigating the ex post spillover effect of positive, negative, and no recognition from one task to another. Contrary to Bradler et al. (2013), who examined the effect of receiving an award on subsequent performance in the same task, the second task in this study was unrelated to the first task. In addition, whereas Neckermann et al. (2014) examined the effect of awards on subsequent performance in another task that did not qualify for an award, the present study included a control group and individuals in both the experimental and the control group could receive symbolic recognition for the second task. Furthermore, while these two studies only addressed the productivity spillover effects of positive awards, I additionally examined how negative recognition influences subsequent task performance.

Negative recognition signals that there is a need for improvement, but it is also something that most individuals hope to avoid. The performance consequences of negative recognition could follow either of two directions. On the one hand, as "losses loom larger than gains" (Kahneman and Tversky 1979), the recipients of negative recognition should invest more effort in preventing future negative recognition. On the other hand, negative recognition is often combined with a decline in self-esteem and status which may result in a performance decrease (Auriol and Renault 2008, Frey and Neckermann 2008, Barankay 2011). Since the effect of negative symbolic recognition on motivation has rarely been studied, it is worth examining whether it leads to an increase or a decrease in performance.

The results of this study show that the response to different symbolic recognition types is heterogeneous across genders. Compared with the students in the control group, the female non-recipients as well as the females who received a smiley-sticker in the estimation task significantly improved in the subsequent flyer-cutting task. However, I do not find a spillover effect of the different recognition types on males' performance.

The paper proceeds as follows. In Section 4.2, I review the related literature. In Section 4.3, I derive the conjectures on the effects of different forms of recognition. Section 4.4 explains the experimental design and procedure of the experiment. In Section 4.5, I present the results. Finally, Section 4.6 discusses the findings of the study and concludes.

4.2 Related literature

This study contributes to three strands of literature. First, it is related to the literature on symbolic recognition and relative performance feedback, second, it discusses the possible underlying mechanisms, and third, it contributes to the literature on gender differences in relative performance settings.

4.2.1 Symbolic recognition and relative performance feedback

The standard economic literature postulates that monetary incentives are crucial for motivating people to work hard (Prendergast 1999, Lazear 2000). However, a recent substantial body of evidence challenges this view and shows that recognition is an important source of individuals' motivation and significantly affects performance (e.g., Herzberg 1966, Ellingsen and Johannesson 2007, Frey and Neckermann 2008, Azmat and Iriberri 2010, Kosfeld and Neckermann 2011, Barankay 2012, Kuhnen and Tymula 2012). Since people derive utility from praise, non-monetary recognition programs are often used to motivate individuals. In addition to reflecting recognition, these programs provide feedback on relative performance. Furthermore, symbolic recognition is associated with low costs and reduces the optimal level of monetary incentives (Besley and Ghatak 2008, Frey and Neckermann 2008, Blanes i Vidal and Nossol 2011). The effectiveness of recognition and feedback is accredited in a meta-analysis by Stajkovic and Luthans (2003). Several studies show that ex ante pronounced recognition programs have a positive effect on performance and work behavior (Markham et al. 2002, Kosfeld and Neckermann 2011, Neckermann and Frey 2013). Kosfeld and Neckermann (2011) found that performance increased, on average, by 12% when the participants could win an award. Furthermore, they found that high ability individuals were more motivated, while low ability persons were less stimulated by awards. In a quasi-experimental setting, Markham et al. (2002) examined the effect of a public recognition program and showed that awards reduced absenteeism by more than 50%. Neckermann and Frey (2013) also showed that hypothetical awards have a significant positive effect on the willingness to contribute to a public good.

Apart from the ex ante incentive effect of awards, research has shown that they can also have an ex post effect on subsequent performance (Bradler et al. 2013, Neckermann et al. 2014). While Neckermann et al. (2014) found an ex post effect of awards on subsequent task performance even when this unrelated task did not qualify for an award, Bradler et al. (2013) provided evidence that unannounced recognition significantly improved subsequent performance in the same task. Both studies suggested that performance increases are mainly driven by strong positive effects of non-recipients.

Despite the literature suggesting that awards have a positive impact on performance, several studies have shown that such programs do not always increase employees' motivation but instead reduce their effort. In the field study by Gubler et al. (2013), awards were implemented as an incentive to increase attendance. However, they found that the award program had a significant negative impact on the net productivity. In addition, Malmendier and Tate (2009) revealed that CEOs had a significant performance decline after receiving an award.

Besides awards, which provide positive performance feedback in combination with symbolic recognition, several studies have suggested that relative performance feedback in itself is decisive for changes in performance. For example, Barankay (2012) showed that removing feedback on the relative performance of furniture salespeople increased their sales performance by 11%. In contrast, Kuhnen and Tymula (2012) found that people exerted more effort when they were informed that they were likely to receive feedback. Whereas Barankay (2012) suggested that effort losses were driven by males who achieved a lower rank than they expected, Kuhnen and Tymula (2012) showed that when the rank was worse than expected, individuals increased their effort. When the feedback was better than expected, Kuhnen and Tymula (2012) found that individuals decreased their output, while in the field experiment by Barankay (2012), no effect could be observed. Other studies providing evidence that information on relative performance leads to higher motivation and effort were conducted by Blanes i Vidal and Nossol (2011), Tran and Zeckhauser (2012), and Azmat and Iriberri (2010). Eriksson et al. (2009) in turn found no impact of relative performance feedback on employee effort.

As mentioned in the introduction, symbolic recognition and relative performance feedback have long been part of the school system. Analyzing data from a natural field experiment in Spain, Azmat and Iriberri (2010) found that relative performance feedback increased high school students' grades by 5%. Tran and Zeckhauser (2012) also showed that information about ranks increased the performance of Vietnamese students in English tests. Related studies that further indicated that recognition programs have a positive effect on students' test scores were carried out by Levitt et al. (2012), Ashraf et al. (2014), and Bandiera et al. (2009).

4.2.2 Symbolic recognition and the underlying behavioral mechanisms

Although there is plentiful empirical evidence that awards and relative performance feedback influence individuals' motivation, how they do so is a controversial issue in the literature. Research has suggested different behavioral mechanisms that explain why awards affect performance. One possible mechanism is individuals' concern about their self-image (Bénabou and Tirole 2006, Köszegi 2006). Symbolic relative performance feedback may increase employees' motivation by enabling individuals to experience greater self-esteem. Self-esteem reflects the respect that a person has for him- or herself and is linked to an individual's belief that he or she possesses the ability to reach a goal. Maslow (1943) identified self-esteem as one of the major sources of motivation. However, negative feedback might also discourage individuals and reduce their self-esteem (Barankay 2011). Relative performance feedback and awards may further crowd out intrinsic motivation when they are perceived as controlling, because then they reduce the image value of high performance (Deci et al. 2001, Ariely et al. 2009).

Another possible mechanism for the motivational effect of awards is status concerns. There is an increasing number of economic theories and empirical studies suggesting that people care about their relative position, and individuals' motivation is often based on status concerns (Moldovanu et al. 2007, Auriol and Renault 2008, Tran and Zeckhauser 2012, Bhattacharya and Dugar 2012, Auriol et al. 2015). The scarce nature of positive recognition, which leads to a distinction between recipients and non-recipients, might result in a status difference. Research has shown that recipients might have a higher status and exert more effort than non-recipients within the social group (Markham et al. 2002, Besley and Ghatak 2008, Kosfeld and Neckermann 2011). However, many studies have also found that status differences are only enjoyed by those with a high status and disliked by those with a low status. As a result, individuals with a low relative position lose motivation (Auriol and Renault 2008, Frey and Neckermann 2008, Barankay 2012). In line with these findings, Ashraf et al. (2014) found that employer recognition and social visibility had a positive impact on performance, while social comparison led to lower performance, especially among low-ability workers.

Changes in performance can also be explained by reciprocity. As employees might consider an award as a gift, they might feel encouraged to give something back and exert greater effort (Akerlof 1982, Fehr et al. 1998, Fehr and Gächter 1998, Fehr and Gächter 2000). While most studies have focused on the reciprocal behavior in response to higher wages, Kube et al. (2012) suggested that non-monetary gifts are particularly powerful in achieving higher work performance. A related concept is conditional altruism, which implies that awards influence effort when employees perceive the award as a signal that the employer is altruistic and cares about them (Levine 1998, Dur 2009).

Another possible mechanism that explains why awards affect the subsequent performance is based on conformity preferences. Bernheim (1994) showed that individuals derived utility from conforming to a social norm because they recognized that a small departure goes along with a loss in popularity. As awards provide feedback on the relative position of a person within the social group, conformity preference can have different implications for recipients and non-recipients (Chen et al. 2010, Bradler et al. 2013, Hoogveld and Zubanov 2014). Chen et al. (2010) investigated the effect of rank information on the number of online ratings. They found that failing to meet the median user's rating increased effort and exceeding the median decreased effort. In the experiment by Bradler et al. (2013), a greeting card was handed out to the best performers in a data-entry task. The authors found that non-recipients were responsible for the performance increase as they tried to improve by reaching the norm. Berger and Pope (2011) suggested that individuals who were only slightly behind others increased their effort and thus were more likely to win. However, they further provided evidence that individuals who were far behind the norm were more prone to give up and reduce their effort levels.

4.2.3 Gender differences and relative performance feedback

Gender differences may clarify the question of why some people care more about their relative position than others. The results relating to gender differences vary across studies. Several papers have documented that women react less favorably to competition than men and that relative to women, men's performance improves under competition (Gneezy et al. 2003, Gneezy and Rustichini 2004, Niederle and Vesterlund 2007). Kuhnen and Tymula (2012) also showed that men performed significantly better than women when information about relative performance was made available. Their results further indicated that men expected to be ranked higher than women, suggesting that men possess greater self-esteem. In line with this finding, Huberman et al. (2004) revealed that men demonstrated a more pronounced status-seeking behavior than women. Related to prior research on gender differences in attitudes toward competition, Barankay (2012) provided evidence that rank incentives especially influenced the performance of men while the behavior of women was unaffected. In contrast to the studies mentioned above, he showed that relative performance feedback led to effort losses, which were driven by males, who achieved a lower rank than they expected. Moreover, the literature has suggested that gender differences under competitive incentives can be sensitive to the type of task. For a male-stereotypical task, Günther et al. (2010) replicated the results of Gneezy et al. (2003) and showed that with competitive incentives men outperformed women. However, in a female-oriented task, women reacted more strongly to competitiveness than men. Shurchkov (2012) also examined gender differences in attitudes toward competition in stereotypical tasks and found that under high pressure men significantly outperformed women in a math task, classified as a male-oriented task, while under low pressure women outperformed men in a verbal task, representing a female-oriented task. Interestingly, Cotton et al. (2013) used a math task and found no evidence for males' advantage after the first period in a contest. In their experiment, women even outperformed men in later periods. In contrast to the previously mentioned studies, Dreber et al. (2011) found no gender differences in the effect of competition for either male-oriented or female-oriented tasks. Furthermore, Dreber et al. (2014) looked at gender differences among adolescents in competitive environments. Their data suggested that no gender difference in performance exists under competitive incentives.[1]

4.3 Conjectures

The primary interest of this study lies in analyzing whether symbolic recognition in one task affects the performance in an unrelated subsequent task and, if so, whether the effect differs with positive, negative, and no recognition. The impact of positive recognition on subsequent performance is not obvious. On the one hand, the subsequent task performance may decrease, because positively recognized individuals are satisfied with their past success and thus consider further effort unnecessary. On the other hand, recognition may raise employees' motivation and performance even in a subsequent unrelated task, because individuals want to live up to the experience of greater self-esteem (Bénabou and Tirole 2006) or they feel inclined to reciprocate (Akerlof 1982, Fehr et al. 1998, Fehr and Gächter 1998, Fehr and Gächter 2000). Further, status concerns might provide an explanation for an increase in the subsequent task performance. Recipients may want to remain in their relatively better position and thus also exert a high level of effort in the second task, whereby they can again obtain symbolic recognition.

[1] Dreber et al. (2014) found no gender gap in the incentive change but showed that male individuals are more likely to self-select into the competitive environment in the male-oriented task. This finding was reconfirmed by Grosse and Reiner (2010), who also indicated that competition entry decisions are driven by gender-oriented tasks.

As the effect of positive recognition is ex ante ambiguous, I postulate the following two conjectures:

Conjecture 1a: *If positive recognition increases self-esteem, status or reciprocal behavior, the subsequent task performance of recipients of positive recognition is higher than that of other individuals.*

Conjecture 1b: *If, however, individuals consider further effort unnecessary and rest on their past success, the subsequent task performance of recipients of positive recognition is lower than that of other individuals.*

The next question is whether there is a performance effect when individuals only receive the information that they are ranked in the middle. As mentioned above, the literature has shown that individuals who receive no award improve their subsequent performance (Bradler et al. 2013, Neckermann et al. 2014). Bradler et al. (2013) identified conformity preference as the key mechanism behind this effect. When learning that their performance is at an average level, individuals tend to increase their effort levels in order to reach the apparent group norm, which in our case is the positive recognition group. An alternative explanation for increased effort provision could be that positive awards serve as a goal that in turn acts as a reference point. Due to loss aversion and diminishing sensitivity, people increase their effort, particularly if they are slightly behind their goal (Berger and Pope 2011). Non-recipients who narrowly missed positive recognition increase their performance to collect a positive award in the following task. A further explanation for higher performance is that status concerns motivate non-recipients to catch up in the second task (Moldovanu et al. 2007). Therefore, the second conjecture is as follows:

Conjecture 2: *Due to conformity preference, reference-based preferences or status concerns, the subsequent task performance of non-recipients is higher than that of other individuals.*

As postulated in the research on relative performance feedback, recipients of negative recognition experience a decline in their self-esteem (Barankay 2011) or a reduction of their status within the social group (Auriol and Renault 2008, Frey and Neckermann 2008, Barankay 2012). Both mechanisms lead to a decrease in motivation and in turn their subsequent task performance could also diminish. Another explanation could evolve from reciprocal preferences (Fehr and Gächter 1998). If reciprocity is the crucial mechanism behind subsequent effort change, recipients of negative recognition will decrease their effort. Alternatively, to reach the gain domain, which in our case is represented by the positive recognition group, individuals with reference-dependent preferences might increase their effort level after receiving negative recognition in the first task. However, due to diminishing sensitivity, individuals who received negative recognition should work less than non-recipients for whom the goal of positive recognition is relatively close. From this argumentation, I derive the following two conjectures:

Conjecture 3a: *If negative recognition leads to a decline in self-esteem, a lower status position or negative reciprocal behavior, the subsequent task performance of recipients of negative recognition is lower than that of other individuals.*

Conjecture 3b: *However, if reference-dependent preferences are the determining mechanism, the subsequent performance of recipients of negative recognition is higher than the performance of individuals who received positive recognition but lower than that of non-recipients.*

Whereas the estimation task is not linked to any gender-specific characteristics, the flyer-cutting is comparable to handicrafts, which in turn are associated with female-stereotyped characteristics like manual speed and coordination (Kimura 1996). Thus, the flyer-cutting task may evoke gender-specific association. Based on these considerations, I suggest that:

Conjecture 4: *While there are no gender-related performance differences in the estimation task, females outperform males in the flyer-cutting task.*

Several studies have shown that the gender differences regarding relative performance feedback depend on the nature of the task (Günther et al. 2010, Grosse and Reiner 2010, Shurchkov 2012). This research suggested when faced with competitive incentives men outperform women in male-oriented tasks, whereas women tend to react equally or are even more sensitive to competitive incentives in female-oriented tasks. Given that cutting flyers is indeed a female-stereotyped task, I predict that:

Conjecture 5: *Female students are more sensitive to the different recognition types than male students.*

4.4 Experimental design and procedure

4.4.1 Experimental design

The experiment was designed to estimate the causal effect of symbolic recognition for a previous task on the performance in a subsequent unrelated task. The experiment consisted of two main tasks. The first task was a simple estimation task from Falk and Zimmermann (2012). Students were shown a picture displaying a bowl filled with peas and were asked to estimate the number of peas inside it.[2] After all the students had written

[2] Instructions, including the picture of the bowl, can be found in Appendix A.

down their estimates, in both treatments the experimenters announced the real number of peas in the bowl which was 3000. In the control treatment, the students continued with the second task. In the experimental treatment, the experimenters announced that the top third would receive a smiley-sticker, the bottom third a frowny-sticker, and the intermediate third nothing other than a message saying that they were ranked in the middle. After the stickers had been distributed, the experimenters asked the participants to put the sticker on their sweatshirt. This procedure ensured that in the experimental group the top and the bottom performers' effort was symbolically recognized in a publicly observable manner.

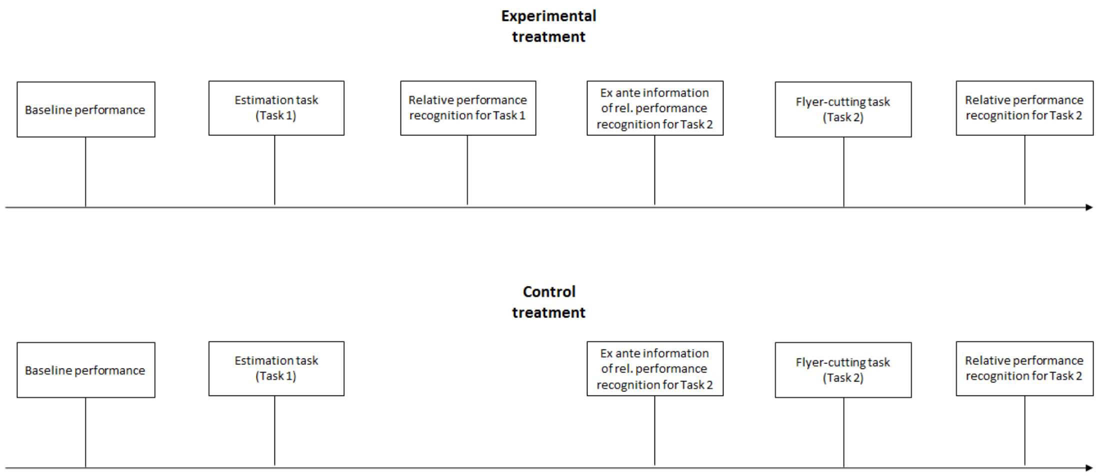
In the second main task, the students were asked to cut out flyers from DIN A4 sheets. Each workstation was provided with scissors and a stack of sheets printed with the flyers.[3] On each sheet, four flyers with varying difficulty levels were printed in a random order. The task was to cut along a black line with a tolerance space of four millimeters shaded in gray. Due to differences in the difficulty level of the motifs, the students were not allowed to start a new sheet before they had cut out all four flyers from the previous sheet. This mechanism prevented students from working just on the easy motifs. The flyers promoted an upcoming concert by a university student orchestra. The participants knew that all the flyers that had been cut out correctly would be used for advertisement. The cutting task offered several advantages. First and most importantly, compared with other real effort tasks (e.g., counting numbers, mathematical problems, transcribing), the flyer-cutting task introduced in this study was not useless, since flyers served as a real-world promotion campaign. Second, it required no special knowledge or cognitive abilities, and it was easy to explain. Third, the task assured a quantity- and quality-based performance measurement. Quantity is measured in the form of the number of flyers cut out and quality as the number of motifs that were cut out within the gray shaded area. Fourth, the learning possibilities were trivial, as the school students, who were at the secondary level of education, already knew before the experiment how to cut out motifs with scissors.

Before the participants started the second task, the students received information on the non-monetary incentive scheme. The experimenters in both treatments pronounced that the third of students with the largest number of correctly cut out flyers would receive a smiley-sticker, while the worst-performing third would receive a frowny-sticker, and the rest would just obtain a message notifying them that their performance was average. As this second recognition system was introduced in both the control and the experimental group, it allows me to isolate the ex post spillover effect of relative symbolic recognition from the ex ante incentive effect. The students had 15 minutes to work on the task. Nobody completed all the sheets provided within the 15 minutes. The participants were

[3] One example of the sheets is shown in Appendix A.

asked to put the cut out flyers into a non-transparent bag marked with their identification number. This procedure minimized peer effects as it prevented the participants from comparing their work with each other (Mas and Moretti 2009). After the students had completed the flyer-cutting task, they received the corresponding recognition incentive for the second task.

FIGURE 4.1: Timeline of the experiment



Prior to the two main tasks, I introduced a baseline stage, which was, apart from the time provided and the non-monetary incentive, identical to the cutting task. The students had five minutes to cut out motifs. I conducted this baseline stage so that the participants could become used to the task. In addition, it provided me with a baseline measure of their ability in a non-competitive and non-incentivized setting. The participants were also asked to fill in a questionnaire, which included basic demographic information like age, gender, and level of education as well as a question regarding whether the students were right- or left-handed because the scissors provided were appropriate for right-handers. The timeline for both treatments is illustrated in Figure 4.1.

4.4.2 Procedure

In total, 138 students took part in the experiment. The participants were students of the seventh, eighth, and ninth graded from a secondary school in Switzerland. Although participation was voluntary, all the students in attendance took part in the experiment. This is most likely because the experiment took place during school time and in the students' classrooms. The experiment was conducted in May 2014 and lasted for approximately 1 hour. A total of 103 students participated in the experimental treatment and 35 students took part in the control treatment. This ensured that about one fourth of the participants received a smiley-sticker for the estimation task and another fourth a frowny-sticker, while another fourth received a message saying that their performance was average. The remaining fourth of the participants took part in the control treatment in which neither a recognition incentive nor feedback was provided for the estimation task. Because of space constraints, the experimental group was divided into three sub-groups. The experiment was conducted simultaneously in four classrooms. This allowed me to rule out the possibility that students would hear about the treatments before they actually took part in the experiment. In the experimental group, in each classroom, the top third received positive symbolic recognition, the bottom third negative symbolic recognition, and the intermediate third only the information that they had performed averagely. Since all four experimenters followed a strict protocol, the procedure for the experimental sub-group was exactly the same. After a short introduction, all the participants were randomly assigned to treatments and classrooms. Before the students were seated in their assigned classrooms, the experimenters mentioned that all the participants would receive a fixed payment of 8 CHF at the end of the experiment and that the payment was independent of their performance in the experiment. The payment of 8 CHF corresponds to the average daily pocket money of a Swiss student at that age. In the classrooms, workstations with all the necessary material were arranged in a way that ensured sufficient space for each student to feel unobserved and to work undisturbed. In both treatments, the participants received written instructions for each task separately and were also asked to answer control questions for the cutting task.

The participants' average age was 14 years, with a standard deviation of 1.2 years, and 47% of the students were female.[4] In the experimental group, 50 girls and 52 boys participated and the control group consisted of 15 girls and 20 boys. While 63 students attended the lower secondary school, 70 participants were enrolled in the upper secondary school.[5]

4.5 Results

The results section proceeds as follows. First, I examine the effect of different recognition types on subsequent performance in an unrelated task, and second, I compare the spillover effect of different recognition types across genders.

[4]One person did not reveal his or her sex.

[5]In Swiss secondary schools, students are separated according to their capabilities. Students with better grades are assigned to upper secondary level, whereas students who aspire to an apprenticeship are allocated to the so-called lower secondary school. Students from both types participated in this experiment. Three persons did not reveal the school type.

4.5.1 Performance spillover effects by recognition type

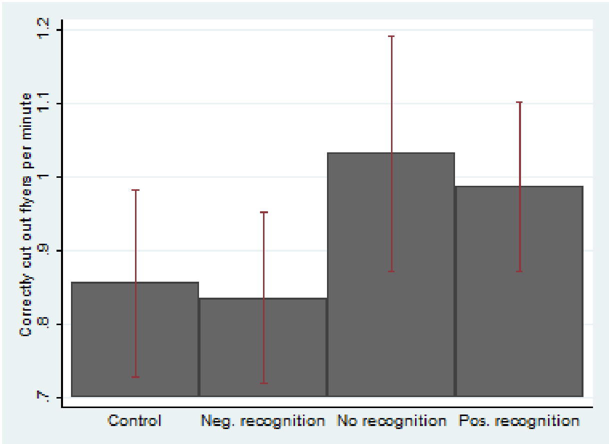
Recognition groups are classified according to the recognition type received in the estimation task. While in the control treatment nobody received performance feedback for the estimation task, in the experimental treatment the top third was awarded a smiley-sticker (n=37), the bottom third received a frowny-sticker (n=32), and the intermediate third received only the information that their performance was average (n=34). If two or more students had the same threshold rank, they were assigned to the higher recognition incentive. As the baseline stage and the flyer-cutting stage differed in length, the performance was measured in terms of productivity. The performance measurement is the number of correctly cut out flyers per minute.[6]

Figure 4.2 presents the average performance in the flyer-cutting task for the different recognition groups and the control group. It shows that the participants who did not receive symbolic recognition and the students who received positive recognition in the estimation task cut out more flyers correctly than the individuals in the control treatment or the students who received negative recognition. On average, the students who did not receive recognition in the estimation task cut out 1.03 flyers correctly per minute, whereas the recipients of a smiley-sticker achieved 0.99 correctly cut out flyers per minute, the individuals in the control group produced 0.86 correct flyers, and the individuals who received a frowny-sticker cut out 0.84 flyers correctly per minute. A Mann-Whitney test shows that the individuals who received no recognition cut out significantly more flyers correctly per minute than the students from the control group (p=0.045).[7] The differences between the control group and either the positive recognition group (p=0.159; Mann-Whitney test) or the negative recognition group (p=0.920; Mann-Whitney test) are not significant. Comparing the average performance across recognition groups, a Mann-Whitney test reveals that the difference between the negative recognition group and the no recognition group is statistically significant (p=0.042). However, the average performance did not significantly differ either between the positive recognition group and the negative recognition group (p=0.126; Mann-Whitney test) or between the positive recognition group and the no recognition group (p=0.522; Mann-Whitney test).

[6]Since relative recognition in the flyer-cutting stage was linked to the number of correctly cut out flyers, the analysis focuses on this performance measurement throughout the paper. The corresponding analysis dealing with the total number of cut-out flyers as performance measurement can be found in Appendix B.

[7]All statistical tests are two-sided.

Figure 4.2: Correctly cut out flyers per minute by recognition type



The figure shows means and error bars.

Table 4.1 reports the descriptive statistics by treatment and recognition group. It shows that the groups are not balanced with respect to their baseline performance. The students in the negative and in the positive recognition group cut out significantly more flyers correctly in the baseline stage than the individuals in the control group (p=0.077 or, respectively, p=0.010; Mann-Whitney test). Regarding the other covariates, the recognition groups and the control groups can be considered as balanced.[8]

As the baseline performance differs across recognition groups, a simple comparison of correctly cut out flyers in the second task is not sufficient. Thus, to provide precise estimates of the spillover effect of different recognition types from one task to another unrelated task, I applied a regression model in which I control for the baseline performance. I run regressions of the following form:

$$y_{i,t=2} = \beta_0 + \beta_1 NegR_i + \beta_2 NoR_i + \beta_3 PosR_i + \beta_4 y_{i,t=1} + \beta_6 \mathbf{X}_i + \epsilon_i \tag{4.1}$$

where $y_{i,t=2}$ is the number of correctly cut out flyers per minute in the second task and $y_{i,t=1}$ represents the baseline performance. I consider three different ranks in the experimental group: the top third ($PosR_i$), who received a smiley-sticker, the intermediate third (NoR_i), who had no symbolic recognition but the information that

[8]There are statistically significant differences regarding the estimation deviation between the negative recognition group and the control group ($p<0.001$; Mann-Whitney test) as well as between the positive recognition group and the control group ($p<0.001$; Mann-Whitney test). However, these differences in the accuracy of estimation are caused by the experimental manipulation itself, and when comparing the average estimation deviation between the complete experimental group and the control group, the differences are not significant.

their performance was average, and the bottom third ($NegR_i$), who received a frowny-sticker. The vector $\mathbf{X}_i$ represents other control variables. In all the specifications, the standard errors are clustered on individual levels.

TABLE 4.1: Summary statistics by recognition type

	Control	Experimental	Neg. recognition	No recognition	Pos. recognition
N	35	103	31	34	37
Female	0.43	0.49	0.55	0.53	0.41
	(0.50)	(0.50)	(0.51)	(0.51)	(0.50)
Age	13.89	13.68	13.52	13.71	13.78
	(1.18)	(1.25)	(1.15)	(1.22)	(1.38)
Upper secondary school	0.53	0.53	0.58	0.49	0.64
	(0.51)	(0.50)	(0.50)	(0.49)	(0.49)
Right-hander	0.89	0.86	0.90	0.79	0.89
	(0.32)	(0.35)	(0.30)	(0.41)	(0.31)
Estimation deviation	2183.26	4.761.27	10759.03***	2433.82	1712.76***
	(650.96)	(22984.6)	(40967.68)	(145.89)	(481.65)
Flyers baseline	0.77	0.77	0.74	0.81	0.77
	(0.31)	(0.28)	(0.25)	(0.29)	(0.29)
Correct flyers baseline	0.39	0.56**	0.54*	0.52	0.61**
	(0.30)	(0.34)	(0.31)	(0.37)	(0.34)
Flyers main task	1.06	1.09	0.99	1.17	1.11
	(0.43)	(0.38)	(0.29)	(0.44)	(0.36)
Correct flyers main task	0.86	0.95	0.84	1.03**	0.99
	(0.37)	(0.39)	(0.39)	(0.46)	(0.34)

The table reports the means for each recognition group. The standard deviations are displayed in parentheses. The Mann-Whitney test is used for numerical data and the Chi-squared test for categorical data. The significance levels indicate a difference of means compared with the control group. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 4.2: Effect of different recognition types on correctly cut out flyers

	1	2	3
Negative recognition	-0.020	-0.075	-0.036
	(0.085)	(0.078)	(0.080)
No recognition	0.176*	0.126	0.147
	(0.101)	(0.098)	(0.094)
Positive recognition	0.132	0.051	0.090
	(0.085)	(0.081)	(0.080)
Baseline performance		0.384***	0.304***
		(0.100)	(0.104)
Male			-0.139**
			(0.064)
Upper secondary school			-0.006
			(0.063)
Right-handed			0.010
			(0.100)
Age			0.060**
			(0.028)
Constant	0.855***	0.704***	-0.030
	(0.063)	(0.081)	(0.350)
N	138	138	135
R^2	0.047	0.156	0.216

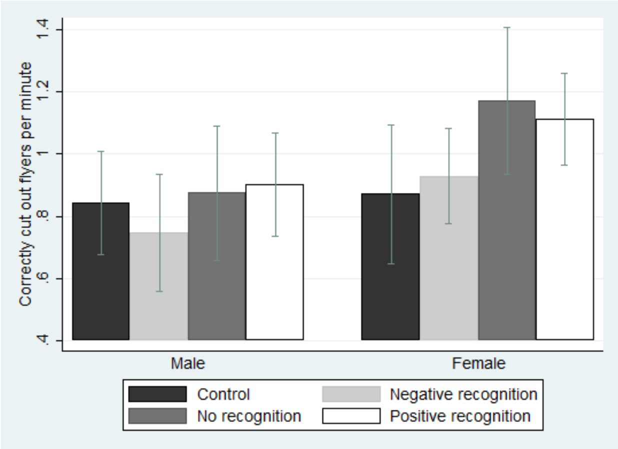
The table shows the OLS estimates. The robust standard errors clustered on individual levels are in parentheses. The sample sizes differ because three students did not reveal the school type and one of them further did not report his or her sex. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.2 presents the results of the OLS regression analysis. Dummies for negative, no, and positive recognition are included in all the specifications. The omitted category is the control treatment. In specification 1, the number of correctly cut out flyers in the second task is regressed on the different recognition types. In specification 2, I additionally control for the baseline performance, and in specification 3, I further add gender, age, school type, and whether the students are right- or left-handed as control variables. The results of specification 1 confirm the findings of the descriptive analysis and show that the non-recipients in the experimental group performed significantly better than the individuals assigned to the control treatment. However, when controlling for the baseline performance, this effect diminishes and is no longer significant. Not surprisingly, the results of specifications 2 and 3 show that the baseline performance significantly influences the productivity in the main flyer-cutting task. Specification 3 also indicates that the productivity significantly increases with age. In addition, the regression analysis reveals that males cut out significantly fewer flyers correctly than females. Since boys cut out fewer flyers correctly than girls not only in the main stage but also in the baseline stage (p=0.003 or, respectively, p=0.098; Mann–Whitney test), flyer cutting can be classified as a female-oriented task. Moreover, a Mann–Whitney test shows that the guesses in the estimation task did not differ among genders (p=0.197). These results are in line with conjecture 4. Since the literature has suggested that gender differences are present in the response to relative performance feedback especially in stereotyped tasks (e.g., Gneezy et al. 2003, Gneezy and Rustichini 2004, Niederle and Vesterlund 2011), I consider the interaction effects of different recognition types and genders in the Section 4.5.2.

4.5.2 Recognition effects by gender

I extend the literature on gender differences by examining whether females and males react differently to relative symbolic recognition. To investigate whether the spillover effect of symbolic recognition varies across genders, I consider the interaction effects between recognition type and gender. Figure 4.3 shows the mean performance in the flyer-cutting task by recognition type and gender. It illustrates that, on average, the girls who received no symbolic recognition cut out 1.17 flyers correctly per minute, the female students who received a smiley-sticker 1.11 flyers, the females who had a frowny-sticker 0.93 flyers, and the girls in the control treatment completed 0.87 correctly cut out flyers per minute. The Mann-Whitney tests show that the female non-recipients and the females who received positive recognition cut out significantly more flyers correctly than the female students in the control group (p=0.010 or, respectively, p=0.046). The girls who received no recognition in the estimation task also performed significantly better than the female students in the negative recognition group (p=0.041; Mann-Whitney test). However, the mean performance does not significantly differ either between the female students who received positive recognition and the females who received negative recognition (p=0.191; Mann-Whitney test) or between the female students in the no recognition group and the females in the positive recognition group (p=0.425; Mann-Whitney test). The male students who received a smiley-sticker, on average, cut out 0.90 flyers correctly per minute, the male non-recipients achieved 0.88 flyers per minute, in the control group they completed 0.84 flyers, and those who received a frowny-sticker produced, on average, 0.75 correctly cut out flyers per minute. However, compared with the control group, the male students did not significantly change their performance in response to previous recognition (for the negative recognition group p=0.409, for the no recognition group p=0.873, for the positive recognition group p=0.869; Mann-Whitney test). Furthermore, even though Figure 4.3 indicates that the male students performed worse after receiving a frowny-sticker than after receiving a smiley sticker or not being recognized, the results are not significant (p=0.368 or, respectively, p=0.847; Mann-Whitney test).

FIGURE 4.3: Correctly cut out flyers per minute by recognition type and gender



The figure shows means and error bars.

Table 4.3 reveals the descriptive statistics by treatment and recognition type for girls and boys separately. It shows that there are statistically significant differences between the recognition groups and the control group in some of the covariates. The share of female upper secondary school students was slightly higher in the positive recognition group than in the control group. In addition, the male students in the negative recognition group were on average younger than those in the control group. Furthermore, there are significant differences regarding the estimation deviation between the boys in the control group and those in the experimental group.[9] More critically, the baseline performance of the boys was unbalanced.

[9]As in section 4.5.1, due to the experimental manipulation, there are significant differences between the control group and the recognition groups regarding estimation deviations.

Table 4.3: Summary statistics by recognition type and gender

	Control	Experimental	Neg. recognition	No recognition	Pos. recognition
Female					
N	15	50	17	18	15
Age	13.80	13.80	13.76	13.56	13.53
	(1.42)	(1.42)	(1.30)	(1.38)	(1.30)
Upper secondary school	0.40	0.50	0.47	0.33	0.73*
	(0.51)	(0.50)	(0.51)	(0.49)	(0.46)
Right-hander	0.93	0.84	0.82	0.78	0.93
	(0.26)	(0.37)	(0.39)	(0.43)	(0.26)
Estimation deviation	2473.93	7073.90	16,861.94**	2,411.28**	1,575.93***
	(491.40)	(32913.93)	(56238.84)	(157.64)	(550.75)
Flyers baseline	0.84	0.84	0.78	0.89	0.87
	(0.34)	(0.30)	(0.26)	(0.31)	(0.34)
Correct flyers baseline	0.51	0.64	0.56	0.59	0.77*
	(0.35)	(0.36)	(0.34)	(0.34)	(0.38)
Flyers main task	1.03	1.17*	1.06	1.25**	1.20
	(0.53)	(0.41)	(0.33)	(0.53)	(0.30)
Correct flyers main task	0.87	1.07**	0.93	1.17***	1.11**
	(0.40)	(0.37)	(0.30)	(0.47)	(0.27)
Male					
N	20	52	14	16	22
Age	13.95	13.73	13.21**	13.88	13.95
	(1.00)	(1.21)	(0.89)	(1.02)	(1.43)
Upper secondary school	0.63	0.57	0.71	0.44	0.57
	(0.50)	(0.50)	(0.47)	(0.51)	(0.51)
Right-hander	0.85	0.88	1.00	0.81	0.86
	(0.37)	(0.32)	(0.00)	(0.40)	(0.35)
Estimation deviation	1,965.25	2,573.38*	3,909.71***	2,459.19***	1,806.05
	(680.90)	(1,497.00)	(2,357.23)	(131.78)	(415.91)
Flyers baseline	0.72	0.71	0.71	0.71	0.71
	(0.29)	(0.23)	(0.22)	(0.25)	(0.23)
Correct flyers baseline	0.31	0.49**	0.53**	0.45	0.49**
	(0.24)	(0.31)	(0.29)	(0.40)	(0.25)
Flyers main task	1.08	1.02	0.92	1.08	1.05
	(0.35)	(0.33)	(0.22)	(0.30)	(0.40)
Correct flyers main task	0.84	0.85	0.75	0.88	0.90
	(0.36)	(0.37)	(0.33)	(0.40)	(0.37)

The table reports the means for each recognition group. The standard deviations are displayed in parentheses. The Mann-Whitney test is used for numerical data and the Chi-squared test for categorical data. The significance levels indicate a difference in means (compared with the control group). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The question of how the spillover effect of different recognition types vary by gender is also examined by an OLS regression analysis. Table 4.4 displays the OLS estimates. All the specifications include dummies for the received recognition type, gender, and interactions between the recognition type dummies and gender. The participants in the control treatment serve as the reference group. Specification 2 also controls for the baseline performance and the interaction between baseline performance and gender. In specification 3, I additionally control for age, school type, and whether students are right- or left-handed. The results confirm the descriptive analysis. Taking specification 3 as a benchmark, the female students who received no recognition and the females who were awarded a smiley-sticker in the estimation task cut out significantly more flyers correctly than the female individuals in the control treatment. The females in the no recognition group increased their performance by 0.30 and the females in the positive recognition group by 0.22 correctly cut out flyers per minute. In contrast, the positive effect of a smiley-sticker as well as of no recognition is offset by negative interaction coefficients when we compare the performance of the males in the experimental group with those in the control group. In addition, the significant interaction between no recognition and being male indicates that boys achieved 0.31 fewer flyer than girls when they received only the information that their performance was average. Furthermore, specification 3 shows that the performance slightly but significantly increased with age.

Overall, the data confirm conjecture 5, which suggests that especially in a female-oriented task girls react more sensitively to relative symbolic recognition feedback than boys. Compared with the female control group, girls significantly increased their subsequent performance after either receiving a smiley-sticker or not being recognized in an unrelated previous task. As stated in conjecture 1a, the spillover effect of positive recognition might be caused by increased self-esteem (Bénabou and Tirole 2006), status concerns (Moldovanu et al. 2007, Auriol and Renault 2008) or positive reciprocal behavior (Akerlof 1982). According to this result, conjecture 1a can be confirmed and conjecture 1b rejected for the female sub-group. The result for female non-recipients is in line with conjecture 2 and can be explained by either reference-dependent or conformity preferences. Since the female non-recipients in the estimation task were only slightly behind the recipients of positive recognition, they might have internalized the smiley-sticker as a reference point and due to loss aversion increased their performance (Berger and Pope 2011). Referring to the explanation based on conformity preferences, the female non-recipients did not comply with the group norm of belonging to the top third and hence they might have felt inclined to improve their performance (Bernheim 1994, Bradler et al. 2013).

Table 4.4: Effect of different recognition types and gender on correctly cut out flyers

	1	2	3
Negative recognition	0.058	0.045	0.053
	(0.126)	(0.124)	(0.121)
No recognition	0.299*	0.280*	0.300**
	(0.152)	(0.152)	(0.140)
Positive recognition	0.240*	0.179	0.216**
	(0.124)	(0.111)	(0.107)
Male	-0.028	-0.061	-0.108
	(0.131)	(0.177)	(0.174)
Negative recognition x male	-0.154	-0.247	-0.185
	(0.173)	(0.159)	(0.167)
No recognition x male	-0.268	-0.316	-0.314*
	(0.199)	(0.194)	(0.189)
Positive recognition x male	-0.180	-0.207	-0.227
	(0.168)	(0.157)	(0.154)
Baseline performance		0.230	0.179
		(0.153)	(0.159)
Baseline performance x male		0.254	0.296
		(0.196)	(0.198)
Age			0.064**
			(0.029)
Upper secondary school			-0.019
			(0.065)
Right-handed			0.010
			(0.099)
Constant	0.871***	0.755***	-0.104
	(0.104)	(0.154)	(0.331)
N	137	137	135
R^2	0.114	0.203	0.247

The table shows the OLS estimates. The robust standard errors clustered on individual levels are in parentheses. The sample sizes differ because three students did not reveal the school type and one of them further did not report his or her sex. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.6 Discussion and conclusion

Symbolic recognition can be a powerful and cost-efficient tool for motivating people. A large body of evidence shows that ex ante mentioned positive recognition increases employees' motivation (e.g., Stajkovic and Luthans 2003, Kosfeld and Neckermann 2011, Bradler et al. 2013). However, little is known about the spillover effect of different recognition types for a previous task on the performance in a subsequent unrelated task. The first evidence shows that unannounced positive recognition increases subsequent performance (Bradler et al. 2013, Neckermann et al. 2014). Compared with Bradler et al. (2013) and Neckermann et al. (2014), I examine not only positive recognition but also negative and no recognition combined with the information that the performance was average. The symbolic recognition program that I introduce in this study is closely related to relative performance feedback, since under both incentive systems individuals receive information about their relative performance and the incentives are not linked to monetary payments. The difference is that I provide relative performance information combined with tangible recognition stickers. Furthermore, symbolic recognition, as used in this experiment, may enhance competition. Since several studies have shown that men react differently from women under competitive incentives (e.g., Gneezy et al. 2003, Gneezy and Rustichini 2004, Günther et al. 2010), the two questions of interest are how symbolic recognition influences the performance in general and whether there is a gender-specific spillover effect.

In this paper, I address these questions by conducting an experiment with secondary school students who had to work on two unrelated tasks. In the first task, they had to estimate the number of peas in a bowl, and in the subsequent task, they had to cut out flyers. In the experimental treatment, the students received unannounced symbolic performance feedback after the estimation task. The top third was rewarded with a smiley-sticker, the bottom third received a frowny-sticker, and the intermediate third did not receive any symbolic recognition. The students in the control treatment received no performance feedback after the estimation task. I find that, compared with the control group, the students who received no symbolic recognition in the estimation task significantly increased their subsequent performance in the flyer-cutting task. However, this effect diminishes and becomes non-significant when I control for the baseline performance. More interestingly, data indicates that females were more sensitive to competitive incentives in the form of symbolic recognition for a previous task than males. The female non-recipients and female students who received a smiley-sticker in the estimation task increased their performance in the subsequent flyer-cutting task. These results remain statistically and economically significant when I control for the baseline performance and demographics. For male students I do not find spillover effects.

According to the results, I suggest that gender differences in competitive environments are context-dependent. More specifically, the task type and the number of tasks are crucial factors for gender-specific reactions to relative performance feedback. Since I identify productivity differences between genders in the baseline as well as in the main stage, the flyer-cutting task can be classified as a female-stereotypical activity. Thus, the results are consistent with the findings of Günther et al. (2010) and Shurchkov (2012), which also showed that females react more strongly to the competitive incentives in female-oriented tasks. In addition, while previous research has mainly focused on monetary competitive incentives, this paper sheds light on a gender-specific spillover effect of different recognition types. Most of the studies that examined gender differences in competitiveness used the same task in a one-shot tournament or, respectively, did not explore the performance spillover effect of competitive incentives. One exception is Cotton et al. (2013), who examined gender-specific behavior in repeated contests. Interestingly, they observed that males outperformed females in the first contests but that women outperformed men in subsequent periods. Compared with their study, I use two unrelated tasks and examine whether the spillover effect of symbolic recognition differs among genders. In addition, by examining the behavior of school students under relative symbolic recognition incentives, this paper contributes to the literature on gender differences in a competitive setting among adolescents. As Dreber et al. (2014) mentioned, finding a suitable incentive scheme and environment for this age group is especially important, since adolescents have to take decisions with long-term consequences. Furthermore, the right incentive system might contribute to reducing the still existing gender wage gap and occupational segregation.

Since flyer cutting is a more female-oriented task, girls reacted significantly more strongly to symbolic recognition than boys. The result that positive recognition increases females' motivation in a subsequent unrelated task can be explained by different behavioral mechanisms. First, after receiving positive feedback, individuals perceive greater self-esteem and want to experience that again in the subsequent task (Bénabou and Tirole 2006). Second, this finding can be explained by positive reciprocity (Kube et al. 2012). The receivers of a smiley-sticker might see the symbolic recognition as a gift and reciprocate by increasing their effort. The result that the female students who did not receive symbolic recognition in the estimation task increased their performance in the subsequent flyer-cutting task is consistent with conformity and reference-dependent preferences. Non-recipients might feel inclined to reach the positive recognition group. This result is also consistent with the findings of Bradler et al. (2013), who showed that non-recipients improved their performance to live up to the award group. Moreover, Hoogveld and Zubanov (2014) argued that conformity preferences are the most likely behavioral mechanism behind an increase in performance after receiving feedback. Additionally, a smiley-sticker might act as a reference point. Consequently, individuals who were below their reference point saw their previous performance as a loss and thus increased their effort. This finding is in line with the results of Berger and Pope (2011), which provided evidence that being slightly behind increases performance.

Although the analysis improves our understanding of the spillover effect of different recognition types, it also raises a number of new questions that should be examined by future research. First, since in this study the experimental as well as the control group received pre-announced symbolic recognition in the flyer-cutting task, the evidence presented does not allow a final assessment of the sole spillover effect of different recognition types. A treatment with unannounced symbolic recognition in the estimation task but without recognition in the flyer-cutting task would be appropriate for isolating the pure spillover effect. Future research should address this issue with different recognition types. Second, it might be worth examining the performance spillover effect of different symbolic recognition types in different gender stereo-typed tasks. Third, although school students are an important and worthwhile sample to study, it would be interesting to test how adults react to different recognition types. Fourth, future research should also be directed towards the examination of the underlying behavioral mechanisms, the role of meaningful tasks, and the long-term effect of non-monetary rewards on subsequent performance.

4.7 References

Akerlof, G. A. (1982). Labor contracts as a partial gift exchange. *Quarterly Journal of Economics*, 97(4):543–569.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–555.

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.

Auriol, E., Friebel, G., and von Bieberstein, F. (2015). The firm as the locus of social comparisons: Implications for the optimal design of promotion systems. *Journal of Economic Behavior and Organization*, forthcoming.

Auriol, E. and Renault, R. (2008). Status and incentives. *RAND Journal of Economics*, 39(1):305–326.

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):435–452.

Bandiera, O., Larcinese, V., and Rasul, I. (2009). Blissful ignorance? Evidence from a natural experiment on the effect of individual feedback on performance. *Policy Research Working Paper Series No. 4122.*

Barankay, I. (2011). Rankings and social tournaments: Evidence from a crowd-sourcing experiment. *University of Pennsylvania*, unpublished manuscript.

Barankay, I. (2012). Rank incentives: Evidence from a randomized workplace experiment. *University of Pennsylvania*, unpublished manuscript.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 95(5):1652–1678.

Berger, J. and Pope, D. (2011). Can losing lead to winning? *Management Science*, 57(5):817–827.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5):841–877.

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

Bhattacharya, H. and Dugar, S. (2012). Status incentives and performance. *Managerial and Decision Economics*, 33(7-8):549–563.

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

Bradler, C., Dur, R., Neckermann, S., and Non, A. (2013). Employee recognition and performance: A field experiment. *ZEW-Centre for European Economic Research Discussion Paper*, 13-017.

Chen, Y., Maxwell, H. F., Konstan, J., and Li, S. X. (2010). Social comparisons and contributions to online communities: A field experiment on MovieLens. *American Economic Review*, 100(4):1358–1398.

Cotton, C., McIntyre, F., and Price, J. (2013). Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior & Organization*, 86:52–66.

Deci, E. L., Koestner, R., and Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1):1–27.

Dreber, A., Essen, E. v., and Ranehill, E. (2011). Outrunning the gender gap - Boys and girls compete equally. *Experimental Economics*, 14(4):567–582.

Dreber, A., Essen, E. v., and Ranehill, E. (2014). Gender and competition in adolescence: Task matters. *Experimental Economics*, 17(1):154–172.

Dur, R. (2009). Gift exchange in the workplace: Money or attention? *Journal of the European Economic Association*, 7(2–3):550–560.

Ellingsen, T. and Johannesson, M. (2007). Paying respect. *Journal of Economic Perspectives*, 21(4):135–150.

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

Falk, A. and Zimmermann, F. (2012). Consistency as a signal of skills. *University of Bonn*, unpublished manuscript.

Fehr, E. and Gächter, S. (1998). Reciprocity and economics: The economic implications of homo reciprocans. *European Economic Review*, 42(3–5):845–859.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–182.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, 42(1):1–34.

Frey, B. S. and Neckermann, S. (2008). Awards. *Journal of Psychology*, 216(4):198–208.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Grosse, N. D. and Reiner, G. (2010). Explaining gender differences in competitiveness: Gender-task stereotypes. *Jena Economic Research Papers 2010-017.*

Gubler, T., Larkin, I., and Pierce, L. (2013). The dirty laundry of employee award programs: Evidence from the field. *Harvard Business School NOM Unit Working Paper 13-069.*

Günther, C., Ekinci, N. A., Schwieren, C., and Strobel, M. (2010). Women can't jump? - An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, 75(3):395–401.

Herzberg, F. I. (1966). *Work and the nature of man.* Staples Press, New York.

Hoogveld, N. and Zubanov, N. (2014). The power of (no) recognition: Experimental evidence from the university classroom. *IZA Discussion Paper 7953.*

Huberman, B. A., Loch, C. H., and Öncüler, A. (2004). Status as a valued resource. *Social Psychology Quarterly*, 67(1):103–114.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2):263–292.

Kimura, D. (1996). Sex, sexual orientation and sex hormones influence human cognitive function. *Current Opinion in Neurobiology*, 6(2):259–263.

Kosfeld, M. and Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. *American Economic Journal: Microeconomics*, 3(3):86–99.

Köszegi, B. (2006). Ego utility, overconfidence and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kube, S., Maréchal, M. A., and Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 102(4):1644–1662.

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

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2012). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *NBER Working Paper 18165*.

Malmendier, U. and Tate, G. (2009). Superstar CEO's. *Quarterly Journal of Economics*, 124(4):1593–1638.

Markham, S. E., Scott, K. Dow, and McKee, G. H. (2002). Recognizing good attendance: A longitudinal, quasi-experimental field study. *Personnel Psychology*, 55(3):639–660.

Mas, A. and Moretti, E. (2009). Peers at work. *American Economic Review*, 99(1):112–145.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4):370–396.

Moldovanu, B., Sela, A., and Shi, X. (2007). Contests for status. *Journal of Political Economy*, 115(2):338–363.

Neckermann, S., Cueni, R., and Frey, B. S. (2014). Awards at work. *Labour Economics*, 31:205–217.

Neckermann, S. and Frey, B. S. (2013). And the winner is…? The motivating power of employee awards. *Journal of Socio-Economics*, 46:66–77.

Nelson, B. (2006). *1001 Ways to Reward your Employees.* Workman Publishing Company, New York, 2nd edition.

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

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1):601–630.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37(1):7–63.

Shurchkov, O. (2012). Under pressure: Gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10(5):1189–1213.

Stajkovic, A. D. and Luthans, F. (2003). Behavioral managment and task performance in organizations: Conceptual background, meta-analysis, and test of alternative models. *Personnel Psychology*, 56(1):155–194.

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

4.8 Appendix A: Experimental instructions

(Original instructions are in German)

General instructions

Welcome to this experiment and thank you very much for your participation. Please read these general instructions carefully:

- The experiment consists of three parts and a short questionnaire.
- The single tasks are explained thoroughly with simple instructions. If you have any questions please raise your hand clearly. We will then come to your place and answer your questions.
- During the experiment you may not use any other devices than those that are mentioned in the following instructions. Please consider that you are not allowed to speak during the experiment. The use of mobile phones, smartphones, tablet-PCs, and so on is forbidden. Interferences lead to exclusion from the experiment.
- As a matter of course, all information is evaluated absolutely confidentially and anonymously.
- Provided that you do not breach these rules, you will receive compensation of 8 CHF.

Thank you very much for your participation in the experiment.

Instructions: Baseline stage

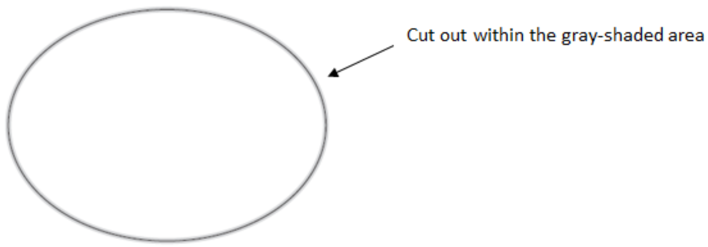
For the first task you have scissors and a stack of A4 sheets with flyers on the desk. The Aarau Students' Orchestra (ASTOR) will play a concert in the church of Buchs next Sunday, on Mother's Day. To have some additional promotion, we kindly ask you to cut out the flyers according to the motifs. This evening, the members of the orchestra will distribute these flyers to the public.

We kindly ask you to cut out as many motifs as possible following the pattern during 5 minutes.

- On the back of the A4-sheet you will see different motifs.
- Cut out the motifs within the marked frame (see the example picture below). Please cut out each motif individually: good quality can only be achieved like this. As we

need the same number of each motif, it is especially important that you always cut out all four motifs on an A4-sheet and only then start the next A4-sheet.

- Place the finished motifs on the plastic plate on your desk. Please consider that only those motifs that have been cut out within the given frame will be counted and distributed in the end.



Control questions Please answer the following control questions by ticking the right answers:

1. Where do the finished motifs need to be placed?
O On the surface of the table O Into the plastic plate O Into the envelope

2. Which motifs will be counted at the end and distributed by the orchestra?
O All the motifs that have been cut out outside the frame
O All the motifs that have been cut out within the frame
O All the motifs that have been cut out

If you have any questions by now, please raise your hand. The experimenter will come to your place. Otherwise please wait until the experimenter calls to start cutting out.

Instructions: Estimation task

Please estimate the number of peas in the pictured bowl. You should estimate the amount of peas as exactly as possible. Enter your answer into the provided gap.

What do you estimate: how many peas are in the pictured bowl? .... peas

In the case that you have any questions concerning this task, please raise your hand. The experimenter will come to your place. After you have made your estimation, please turn over the sheet. The experimenter will collect it afterwards.

[Note: After the estimation task, the experimenters in the experimental groups made the following announcement: *Great, you have successfully finished this task of the experiment. For your estimation we now bestow awards. The top third of the students in this room receives a smiley-sticker (show example) while the bottom third of all the students in this room receives a frowny-sticker (show example). The intermediate third of all the students in this room receives no sticker but a message that they have performed averagely. Your answers are now being evaluated so that we can afterwards assign the smiley- and frowny-stickers. As soon as the stickers have been distributed, we kindly ask you to stick them to your t-shirt for the rest of the experiment.*]

Instructions: Flyer-cutting task, second stage

This is now the last part of the experiment. Again, you can find scissors and a stack of A4-sheets on the desk. As before, the flyers are for the Aarau Students' Orchestra (ASTOR). This evening they will be distributed by members of the orchestra. After reading the instructions carefully, please repeat the task of the first part during the next 15 minutes.

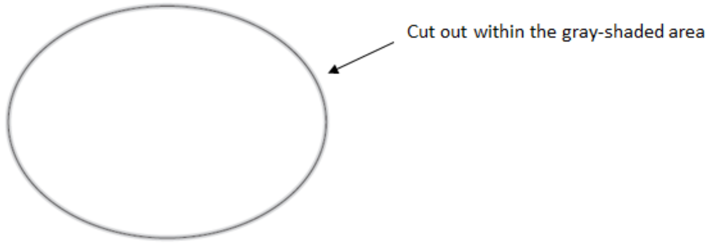
Award

Please consider that you can win an award after completing this task. The top third, meaning the students who cut out the most motifs within the given frame, will receive a smiley-sticker. The bottom third, meaning the students who cut out the fewest motifs within the given frame, will receive a frowny-sticker. The intermediate third will receive no recognition but a message that they have performed averagely.

Task

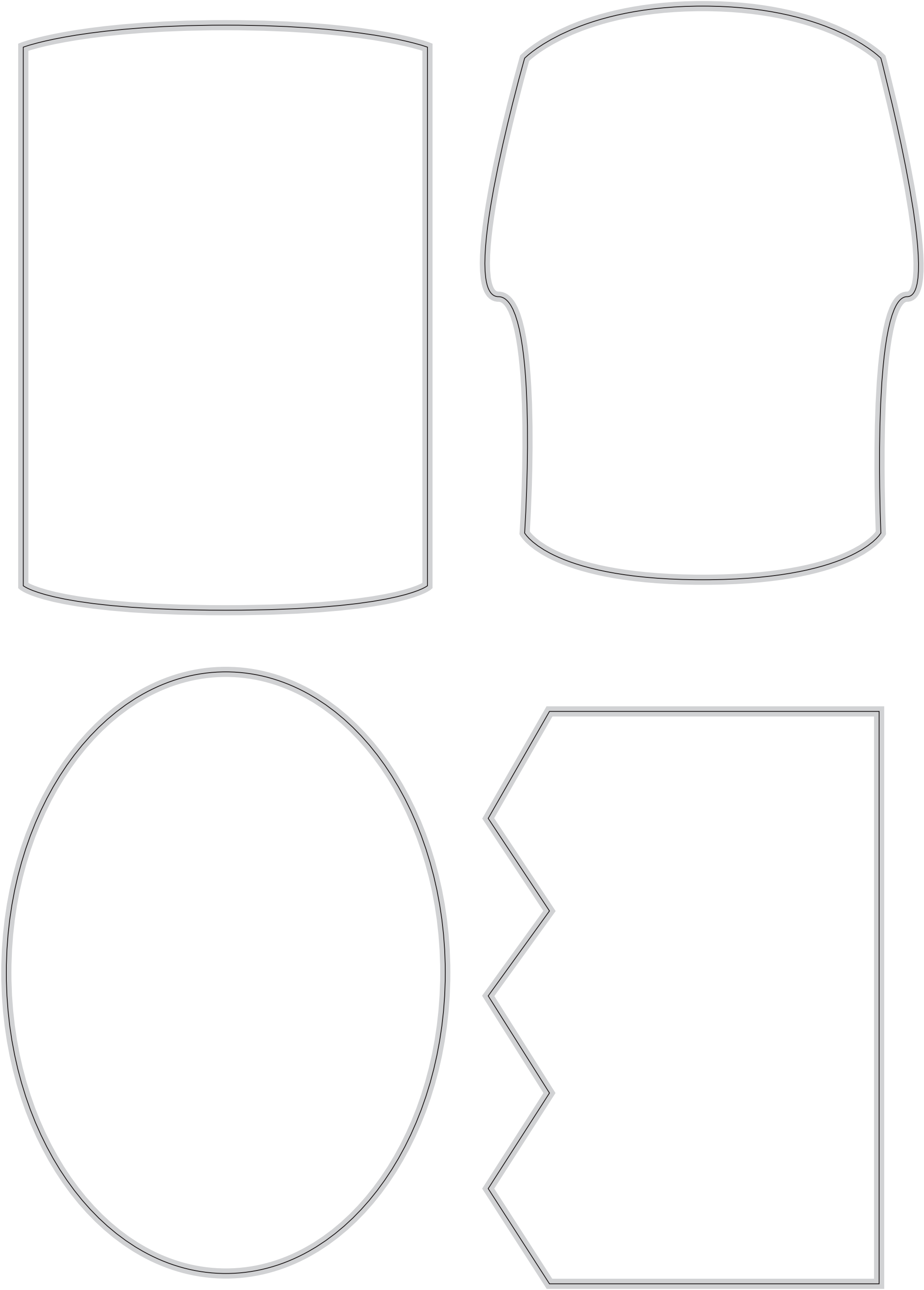
We kindly ask you to cut out as many motifs as possible following the pattern during the next 15 minutes (in exactly the same way as in part 1):

- On the back of the A4-sheet you will see different motifs.
- Cut out the motifs within the marked frame (see the example picture below). Please cut out each motif individually: good quality can only be achieved like this. As we need the same number of each motif, it is especially important that you always cut out all four motifs on an A4-sheet and only then start the next A4-sheet.
- Place the finished motifs on the plastic plate on your desk. Please consider that only those motifs that have been cut out within the given frame will be counted and distributed in the end.



If you have any questions, please raise your hand. The experimenter will come to your place. Otherwise please wait until the experimenter calls to start cutting out.

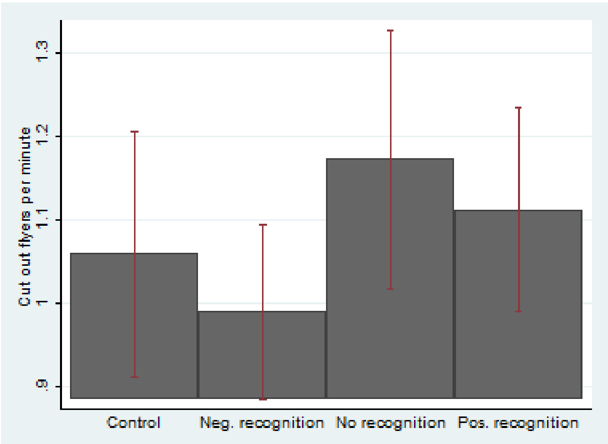
Example of sheets with motifs



4.9 Appendix B: Further results

The following analysis is focused on the number of cut out flyers. Figure 4.4 shows the average number of flyers cut out per minute for each recognition type. It indicates that the students who did not receive recognition in the estimation task cut out the most flyers, whereas the students who received negative recognition cut out the fewest. In specification 1 of Table 4.5, we regress the number of cut out flyers on the recognition types. The participants in the control treatment serve as the reference group. In specification 2, we add the baseline performance, and in specification 3, we further control for demographics. The results reveal that negative, positive, and no recognition for the estimation task have no significant impact on the number of cut out flyers.

FIGURE 4.4: Cut out flyers per minute by recognition type



The figure shows means and error bars.

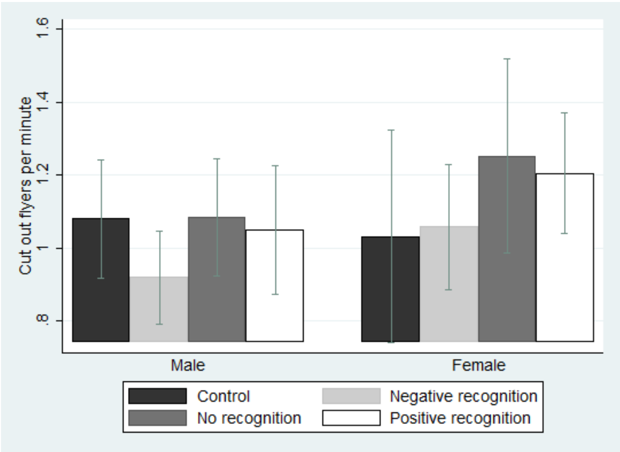
Table 4.5: Effect of different recognition types on cut out flyers

	1	2	3
Negative recognition	-0.069	-0.088	-0.033
	(0.088)	(0.088)	(0.089)
No recognition	0.114	0.097	0.120
	(0.105)	(0.105)	(0.098)
Positive recognition	0.053	0.025	0.073
	(0.094)	(0.093)	(0.088)
Baseline performance		0.129	0.052
		(0.106)	(0.110)
Male			-0.104
			(0.069)
Secondary school			-0.055
			(0.064)
Right-handed			0.025
			(0.097)
Age			0.103***
			(0.027)
Constant	1.059***	1.008***	-0.335
	(0.072)	(0.092)	(0.353)
N	138	138	135
R^2	0.029	0.041	0.162

The table shows the OLS estimates. The robust standard errors clustered on individual levels are in parentheses. The sample sizes differ because three students did not reveal the school type and one of them further did not report his or her sex. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

When taking the gender difference into account, Figure 4.5 reports that the female students who received either no or positive recognition for the estimation task cut out more flyers than the females in the control group or those who received negative recognition. In addition, Figure 4.5 shows that the male students cut out the fewest after receiving negative recognition, whereas those in the control, positive, and no recognition groups reached a higher mean number of cut out flyers. To examine whether females or males cut out more flyers in response to previous symbolic recognition, we additionally regress the number of cut out flyers on the interactions between the symbolic recognition groups and the gender. Table 4.6 shows the estimates. When controlling for the baseline performance and demographics, the results indicate that positive recognition has a slightly significant positive effect on the number of cut out flyers. We do not find any other spillover effects of symbolic recognition on the quantity. Furthermore, specification 3 suggests that age has a significant impact on the number of cut out flyers.

Figure 4.5: Cut out flyers per minute by recognition type and gender



The figure shows means and error bars.

Table 4.6: Effect of different recognition types and gender on cut out flyers

	1	2	3
Negative recognition	0.028	0.023	0.039
	(0.158)	(0.158)	(0.145)
No recognition	0.221	0.214	0.246
	(0.185)	(0.186)	(0.162)
Positive recognition	0.173	0.150	0.224*
	(0.156)	(0.145)	(0.127)
Male	0.049	0.058	0.004
	(0.156)	(0.217)	(0.197)
Negative recognition x male	-0.189	-0.208	-0.127
	(0.185)	(0.184)	(0.185)
No recognition x male	-0.217	-0.226	-0.246
	(0.214)	(0.215)	(0.202)
Positive recognition x male	-0.205	-0.202	-0.264
	(0.194)	(0.187)	(0.176)
Baseline performance		0.086	0.004
		(0.187)	(0.180)
Baseline performance x male		0.024	0.103
		(0.224)	(0.219)
Age			0.107***
			(0.029)
Right-handed			0.027
			(0.101)
Secondary school			-0.070
			(0.067)
Constant	1.031***	0.987***	-0.443
	(0.135)	(0.199)	(0.342)
N	137	137	135
R^2	0.058	0.064	0.179

The table shows the OLS estimates. The robust standard errors clustered on individual levels are in parentheses. The sample sizes differ because three students did not reveal the school type and one of them further did not report his or her sex. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Koautorenschaften sowie alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Arbeit verliehenen Titels berechtigt ist.

Signed: Andrea [signature]

Date: 25.09.2015