UNIVERSITÄT
BERN

Motivation through information: Three field experiments

Inauguraldissertation zur Erlangung der Würde eines

Doctor rerum oeconomicarum

der Wirtschafts- und Sozialwissenschaftlichen Fakultät

der Universität Bern

vorgelegt von

Angela Steffen

Bern, Januar 2019

Die Fakultät hat diese Arbeit am 23. Mai 2019 auf Antrag der beiden Gutachter Prof. Dr. Frauke von Bieberstein und Prof. Dr. Michael Kosfeld als Dissertation angenommen, ohne damit zu den darin ausgesprochenen Auffassungen Stellung nehmen zu wollen.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my first supervisor, Prof. Dr. Frauke von Bieberstein. Frauke gave me the opportunity to start as an external PhD student at her chair, integrated me into her team from the beginning, and provided me with all the guidance and support needed to complete this thesis. I thank Frauke, in particular, for her continuous encouragement and her enthusiasm for experimental research. The motivating collaboration inspired me on a professional and personal level.

I am also very grateful to Prof. Dr. Michael Kosfeld for being my second supervisor and for his valuable time and expertise.

Furthermore, I want to thank all those people with whom I have had the pleasure to work during these projects. I am indebted to the Institute of Tourism at the University of Applied Sciences, Lucerne, for their financial support during my studies. I am also very grateful to my co-authors, Dr. Andrea Essl, Dr. Martin Staehle, and Zita Spillmann, for the enriching teamwork and our stimulating discussions. In particular, I would like to thank Dr. Andrea Essl and Dr. Stefanie Jaussi for their valuable advice throughout the process. I also gratefully acknowledge the helpful assistance provided by Dr. Raquel Andres Martinez, Lucian Hunger, Jonas Collenberg, and Roden Safar. A sincere thanks goes to my colleagues from the Institute of Tourism in Lucerne and the Institute for Human Resource Management and Organization of the University of Bern, who shared with me the challenges and achievements of this thesis.

Finally, I must express my very profound gratitude to my family and friends. I am grateful to my parents, Ester and Daniel Steffen, and to my partner, Adrian Elmiger, for providing me with unfailing moral support and encouragement. This accomplishment would not have been possible without them. Special thanks also to Caspar Van de Ven, Zora Kubinec, and Claudia Bucher for their interest in my research and their valuable comments. My deepest appreciation goes to my sister, Andrea Steffen, who gave me a lot of inspiration and support along these years of study. I dedicate this thesis to her.

Bern, January 2019
Angela Steffen

Contents

List of Figures

List of Tables

Executive Summary

Traditional economic theory promotes tangible rewards and punishments as sufficient drivers of human motivation. Behavioral research, however, shows that humans often depart from the neoclassical assumptions that underlie many incentive programs. Individuals have, for example, limited cognitive resources to make choices that are in their best long-term interests, and these interests are not always of a purely selfish nature. Such insights have important implications for today's management of private and public organizations. Monetary incentives and control mechanisms may backfire by crowding out intrinsic motivation, while merely symbolic or informational interventions can provide effective alternatives. Behavioral economics integrates psychological factors into traditional economic theory to more fully understand human behavior. It thereby suggests new tools to leverage intrinsic motivators and to guide human decision-making in various contexts.

At the core of this research is the idea of motivating individuals through information that does not entail direct monetary consequences. The age of big data and telecommunication technology opens up new avenues for such approaches, as information is increasingly readily available and easy to share. This thesis consists of three studies addressing the question of how organizations can use personal messages to motivate behavioral change. In this context, each study highlights motivation in a different domain. Essay 1 investigates the impact of information at work. Essay 2 deals with behavior in the environmental context, and Essay 3 addresses behavioral change in the social realm.

The common ground of all three essays is the exploration of individual, real-world behavior using field experiments. This methodological approach has the advantage of capturing controlled data in a normally occurring environment. The present field experiments were conducted with three practice partners: a commercial company (Essay 1), a non-profit association (Essay 2), and a charitable organization (Essay 3). These different partner organizations permit an inside view into the diversity of contemporary challenges that behavioral research may help tackle.

Essay 1—a joint work with Frauke von Bieberstein—explores the effect of comparative performance information that is more timely and specific than traditional performance

feedback. By providing real-time performance benchmarks at work, this intervention aims to motivate sales employees of a railway catering company in Switzerland. We find that real-time feedback can significantly increase sales revenues, if it contains information about the average recent performance of co-workers. This effect seems to be persistent over time and is clearly driven by employees at intermediate levels of performance. Workers at the top and at the bottom of the productivity distribution, in contrast, show no significant reactions. In light of the increasing use of timely, 360-degree feedback in practice, our results emphasize the value of real-time feedback for lasting improvements in work productivity.

Essay 2—a joint work with Andrea Essl and Martin Staehle—examines whether and how reminder messages can motivate consumers to take up environmentally friendly habits. This study was conducted with an agricultural association that encourages their customers to return plastic packaging for reuse. Essay 2 demonstrates the beneficial impact of reminders on the return rate of plastic bags, by making this desired behavior more salient. The positive effect does not fade when multiple reminders are applied and also persists, to some extent, beyond the intervention period. The results also provide new evidence for the action-closeness effect of reminders. Reminders are significantly more effective when they catch customers' attention in direct proximity to the recycling decision. This finding indicates how reminder messages may be more successfully implemented in practice.

Essay 3—a joint work with Zita Spillmann—investigates how performance-related information can increase volunteers' prosocial engagement in a German aid organization. More specifically, this study examines whether information about the average workload per volunteer can motivate less-active members to engage more actively in charity. Against the expected conformity effect, we find that this intervention has no significant impact on the number of voluntary service hours. Neither volunteers at above-average levels of engagement, nor volunteers with a below-average commitment, show significant reactions. Additional analyses highlight the risk that already-active volunteers become even more overburdened when the organization intensifies internal communication with its volunteers.

Essay 1: Go beyond (your) average: A field experiment on real-time performance feedback and sales productivity

Angela Steffen, Frauke von Bieberstein [*]

Abstract

Real-time performance feedback is one of the major trends in human resource management. However, insights about the implications of providing ongoing and timely performance information to employees are still scarce. We present the results of a randomized controlled trial involving 164 sales employees of a large railway catering company in Switzerland. In the presence of a relative incentive scheme, we find that real-time information about average performance levels can significantly increase sales productivity. In our setting, we observe a revenue growth of up to 3.9%, which corresponds to over 0.4 million Swiss francs additional revenue per year. This effect is mainly driven by employees performing just below the average productivity level. The top- and poorest-performing workers do not show significant reactions.

Keywords: real-time feedback, field experiment, employee motivation, relative performance feedback

[*]Angela Steffen: angela.steffen@iop.unibe.ch, Institute for Organization and Human Resource Management, University of Bern; Frauke von Bieberstein: frauke.vonbieberstein@iop.unibe.ch, Institute for Organization and Human Resource Management, University of Bern. We thank the management of the partner company for their considerable support in this project. We are grateful to Christian Zehnder, Holger Herz, and the participants of the workshop on feedback and recognition at the Erasmus University Rotterdam 2017, the CUSO workshop 2016, and the internal research seminar at the University of Bern 2017. We also thank Nick Zubanov, Matthias Kräkel, and the participants of the organizational economics session at the Annual Meeting of the Verein für Socialpolitik 2018. All errors are ours.

1.1 Introduction

Organizations are radically changing the way they measure, evaluate, and recognize employee performance. For example, PwC (2015) reports that two-thirds of large companies in the United Kingdom are in the process of adapting their performance management systems. According to Deloitte (2015), 82% of surveyed U.S. companies perceive traditional performance evaluations as not being worth the time. With increasing digitalization, the availability of performance-related information is rapidly expanding. A related trend in the current "performance management revolution" is the shift from year-end appraisals towards a continuous feedback culture with real-time performance reviews (Deloitte 2015, Cappelli and Tavis 2016, The Economist 2016). Goldman Sachs and JP Morgan are just two recent examples of companies where this is happening (Son 2017, Surane 2017).

In this study, we investigate the effect of real-time feedback characterized by the frequent provision of timely performance information. The existing literature has reported positive impacts of real-time feedback in the context of resource consumption (Tiefenbeck et al. 2018), group collaboration (Jung et al. 2010), and logistics processes (Goomas and Ludwig 2007, Ludwig and Goomas 2009). Yet, the precise implications and optimal design of real-time feedback at work are mostly unexplored. In this paper, we compare different types of timely performance information to general performance reviews that are traditionally provided ex post to the assessment period.

We conduct a field experiment in a large Swiss catering enterprise with 164 sales employees who offer drinks and snacks on domestic trains. By randomly assigning subjects to groups, we introduce three experimental treatments where employees frequently receive personal and/or co-worker-related performance information directly at work. In accordance with the relative incentive scheme of the company, the feedback messages either contain an employee's personal average sales revenue over the recent past ("personal info"), the average sales revenue of all employees over the recent past ("social info"), or both ("personal and social info"). This information is given in addition to an aggregated performance summary (i.e., an employee's relative performance across all tasks), which is the basis for relative bonus payments at the end of every month. In contrast to the monthly performance signal, the information provided in our intervention always refers to an employee's current work shift and is updated on a daily basis.

We find that real-time feedback that allows employees to continuously evaluate their performance relative to that of co-workers induces a strong increase in sales productivity. Sales revenues in the "social info" and "personal and social info" treatment groups grow up to 3.3% and 3.9% compared to the control group. Furthermore, these effects seem

to be stable over the intervention period, indicating substantial economic benefits in the longer term. The timely provision of personal performance averages, however, has no significant effect on sales revenues. Additional analyses on employees at different ability levels reveal that the effect of real-time feedback is not uniform. In line with existing evidence on relative performance feedback (Hannan et al. 2008, Casas-Arce and Martínez-Jerez 2009, Bandiera et al. 2013, Delfgaauw et al. 2014), the positive effects in the treatment groups are driven by employees at intermediate levels of performance, particularly by those who usually perform just below average. In contrast, workers at the top and at the bottom of the productivity distribution are not significantly affected. This finding highlights the importance of considering different employee capabilities when introducing new feedback policies in practice.

This paper extends previous studies on feedback frequency (e.g., Kang et al. 2005, Northcraft et al. 2011, Kuhnen and Tymula 2012, Casas-Arce et al. 2017), and feedback immediacy (e.g., Mason and Redmon 2008, Kettle and Häubl 2010, Fajfar et al. 2012) by investigating performance information provided in real time. Employees can therefore immediately adapt their behavior. Our study also broadens existing evidence on real-time feedback effects (Goomas and Ludwig 2007, Jung et al. 2010, Tiefenbeck et al. 2018), as we explore the impact of frequent and timely feedback in a new work setting using real-world information on individual sales performance. Furthermore, our study directly compares real-time performance information to an aggregated performance measure that is periodically revealed through the relative incentive scheme of the company. Insights about effective feedback policies in such settings are important, as relative monetary rewards or workplace tournaments are highly pervasive in practice (McGregor 2006). This paper also contributes to the existing research on relative performance feedback, showing that performance may improve (e.g., Blanes i Vidal and Nossol 2011, Delfgaauw et al. 2013, Blader et al. 2015) or deteriorate (e.g., Barankay 2011b, Bandiera et al. 2013) when employees learn about their relative standing compared to their peers. Our analyses extend these findings by comparing the effect of personal versus peer-related performance information in a real work context.

The remainder of this paper is structured as follows. In Section 1.2, we review existing evidence on timely, comparative performance information and develop our hypotheses. The field setting, experimental design, and field data are set out in Section 1.3. Section 1.4 presents the empirical results. Finally, Section 1.5 discusses the findings and approaches for future research.

1.2 Related literature and hypotheses

In this part, we review the literature relevant to the time- and content-related aspects of our feedback intervention (Sections 1.2.1 and 1.2.2). This literature shows our rationale for the presumed effects on sales productivity. We complete our hypotheses by considering the relationship between feedback effects and ability in Section 1.2.3.

1.2.1 Feedback timing

Feedback is defined as the provision of "information regarding some aspect(s) of one's task performance" (Kluger and DeNisi 1996, p. 255). Such information has been successfully used to increase performance in a variety of organizational settings (Nolan et al. 1999). However, in economic literature and managerial practice, the precise implications of providing frequent and timely feedback at work are mostly unexplored. Existing evidence broadly supports the idea that immediate performance information leads to better performance than delayed feedback (Alavosius and Sulzer-Azaroff 1986, Mason and Redmon 2008, Kettle and Häubl 2010, Fajfar et al. 2012) and that more specific feedback should be more beneficial (Earley et al. 1990, Casas-Arce et al. 2017). However, previous literature on feedback frequency provides mixed results. From a theoretical perspective, our study is related to existing models on interim performance feedback in tournaments (e.g., Yildirim 2005, Aoyagi 2010, Ederer 2010, Goltsman and Mukherjee 2011). This literature highlights that interim feedback creates asymmetries between agents and can affect effort choices before and after its revelation. Whether more frequent feedback increases the principal's payoff depends on the agents' cost of effort functions (Aoyagi 2010, Ederer 2010).

In a similar vein, empirical studies reveal positive and negative outcomes of increasing the frequency of feedback. Chhokar and Wallin (1984), for example, find no effect of more frequent feedback on safety performance. Casas-Arce et al. (2017) show that professionals achieve the highest customer satisfaction scores when they receive detailed but infrequent (i.e., monthly) feedback. As confirmed by Lurie and Swaminathan (2009), this effect arises because workers tend to put too much weight on the most recent information disclosed. On the contrary, So et al. (2013) suggest that more frequent feedback is effective for improving the customer service behaviour of employees at a gas station. Their results indicate small but consistent improvements in service performance when employees receive daily compared to weekly feedback. Kang et al. (2005) find that more frequent feedback produces a higher level of performance than less frequent feedback if individuals receive incentive payments. Similarly, Northcraft et al. (2011) report a positive impact of more frequent and more specific feedback on performance, showing

that the positive effects are accentuated when both characteristics are combined. The findings of Goomas et al. (2011) further indicate that ongoing real-time comparisons with task-specific performance benchmarks (so-called engineered labour standards) have a substantial positive impact on workers' productivity in a warehouse distribution centre (also see Goomas and Ludwig 2007, Ludwig and Goomas 2009).

Based on these results, we would expect a positive impact of real-time feedback on subsequent performance in our setting. In contrast to Casas-Arce et al. (2017) and Lurie and Swaminathan (2009), performance information in our study is not only more frequent but also more specific and dynamic, providing employees with frequent data that is relevant for their *current* work task. This may be considered a beneficial feature of our intervention. Because we test real-time feedback containing different types of performance benchmarks, we now proceed with a literature review on comparative performance information before developing our hypotheses.

1.2.2 Comparative performance feedback

One major explanation for feedback effects is the possibility for self-evaluation. The social psychology literature has repeatedly emphasized that motivation and behavior are regulated by the comparison of personal performance outcomes to an implicit or explicit standard of excellence (Strang et al. 1978, Ilgen et al. 1979, Locke et al. 1981, Bandura and Cervone 1983). Alvero et al. (2001, p. 19) accordingly identifies two types of feedback interventions as being equally popular in the feedback literature: the comparison of an individual's performance to his or her past performance ("temporal comparative information") and the comparison of individual performance with a standard or mean of performance ("social comparative information").

The fact that people are influenced by temporal or social comparative information is documented in various empirical studies. After providing information about the average performance of their peers, individuals, for example, improve their performance in a brainstorming task (Szymanski and Harkins 1987, White et al. 1995), increase curbside recycling (Schultz 1999), and reduce household energy consumption (Schultz et al. 2007). Even the communication of simple, personal performance levels (also defined as knowledge of results) is shown to induce significant performance improvements in different field settings (e.g., Hundal 1969, Kim and Hamner 1976, Crowell et al. 1988, Schultz 1999, Sharma et al. 2016).

Recent economic literature further demonstrates that performance can be effectively enhanced by relative rank feedback, where individuals learn their relative standing compared to their peers (e.g., Blanes i Vidal and Nossol 2011, Kuhnen and Tymula

2012, Tran and Zeckhauser 2012, Azmat and Iriberri 2016). These positive effects occur even when performance is not tied to pecuniary rewards, suggesting that people value relative outcomes per se (also see Klein 1997, Clark et al. 2008). However, several studies report negative (Barankay 2011a,b, Akın and Karagözoğlu 2017) or no effects (Eriksson et al. 2009) of relative performance information. Barankay (2011a) shows that private rank feedback, which is updated on a daily basis, has a significant negative impact on sales performance in a furniture company. On the team level, Bandiera et al. (2013) find that daily histograms on teams' productivity lead to excessive free riding and reduce overall performance if relative productivity is not tied to monetary rewards. Hannan et al. (2008) and Azmat and Iriberri (2016) further underline that the effect of relative performance feedback depends on the incentive scheme, suggesting that it is most beneficial for piece-rate compensation.

In our study, we expect a positive effect of personal and social comparative performance information. Akın and Karagözoğlu (2017) and Eriksson et al. (2009) presume that their negative results are driven by specific features of their designs.[1] In contrast to Barankay (2011a) and Barankay (2011b), we do not conjecture that employees get demoralized by the performance information in our intervention because we do not provide aggregated rank feedback but task-specific, absolute performance benchmarks (see Section 1.3.2). This information is less evaluative and absolute than a ranking order. More importantly, participants in our study can directly react to the feedback messages and enhance their relative performance on the same day. This is in contrast to the B2B context of Barankay (2011a), where sales are "lumpy" because they depend on a few big customers and where salespeople also work on tasks other than selling. Also considering the insights on feedback timing set out in Section 1.2.1, we propose the following:

Hypothesis 1. *Real-time feedback containing personal and/or social performance averages over the recent past increases sales productivity.*

This hypothesis is also supported by existing evidence on peer effects, suggesting that peer monitoring significantly increases work productivity (Falk and Ichino 2006, Mas and Moretti 2009). Peer monitoring basically provides ongoing, co-worker-related performance information, but in contrast to our study and the feedback interventions mentioned above, this information is publicly accessible within teams. The expected positive effects of Hypothesis 1 are presumably further promoted by some specific features of our design. The relative incentive scheme in our setting rewards above-average sales performance with bonus payments and therefore incentivizes productivity increases (see Section 1.3.1).

[1] That is the use of a cognitively demanding task, where feedback is distracting (Akın and Karagözoğlu 2017) and a ceiling effect with subjects already exerting maximum effort given their ability (Eriksson et al. 2009).

Feedback further refers to a task with relatively low cognitive demands and is provided via a computer screen rather than via personal communication (see Section 1.3.2). Both characteristics should reinforce the positive impact of feedback on performance (see Kluger and DeNisi 1996).

Regarding the differential effects of personal versus social performance benchmarks, existing evidence is limited (Moore and Klein 2008, p. 61). Moore and Klein (2008) suggest that information about one's absolute standing may be more influential than social comparative feedback. However, Blader et al. (2015) show in a field experiment with truck drivers that rank information with respect to co-workers leads to better outcomes than information about individual performance only. In our study, we equally expect the effect of social feedback to be stronger. In particular, co-worker-related performance information in our setting is more novel than personal performance feedback. Employees could theoretically track their own performance over the recent past themselves, while the performance of their co-workers is largely unknown (see Section 1.3.1). We further conjecture that the impact of real-time feedback is greatest for the "personal and social info" condition. This direct comparison of personal and co-worker-related performance averages is most closely related to financial incentives in our design (see Section 1.3.1). Previous work confirms that feedback combined with monetary consequences produces more consistent effects than feedback alone (see Alvero et al. 2001). We therefore propose the following:

Hypothesis 2. *The positive effect of real-time feedback is highest when providing personal and social performance averages over the recent past and lowest for information on personal average performance alone.*

1.2.3 Feedback and ability

The varying findings of the literature on relative performance feedback indicate that feedback effects are not homogeneous. This is also supported by the presumed non-linear impact of relative performance information on workers with different capabilities. Referring to the "dynamic incentive effect", existing studies reveal that informing participants about their relative standing during a competition has a hump-shaped effect on performance. Participants who lag far behind and those who are far ahead slack off. However, incentive salience and feedback responsiveness is high for participants at intermediate performance levels (Bartel 2004, Hannan et al. 2008, Casas-Arce and Martínez-Jerez 2009, Delfgaauw et al. 2014). Such feedback effects are also found outside of relative rewards, where from a purely rational perspective the feedback sign should not affect performance (Kuhnen and Tymula 2012). According to social cognitive theory, comparative information, such as personal progress or relative standing,

affect motivation by influencing individuals' perceived capabilities to attain a certain standard (e.g., Bandura and Cervone 1983, Bandura and Jourden 1991, Schunk and Swartz 1993). This leads to a curvilinear relationship between performance–standard discrepancies and an individual's subsequent effort (also see Heckhausen 1977, Feather 1982). Outside of relative incentives, empirical studies confirm the detrimental effects of relative performance feedback on individuals at the bottom (Eriksson et al. 2009, Bandiera et al. 2013) and at the top (Schultz et al. 2007, Fischer 2008) of the performance distribution.

In line with this evidence, we expect that co-worker-related performance information in our setting is more effective for workers at intermediate levels of performance and less effective for lowest- and highest-performing employees. This effect should particularly appear in the "personal and social info" condition, where an employee's relative standing in the reference group becomes most salient. In the "personal info" treatment, the performance–standard discrepancies and the related psychological and monetary consequences for the best- and lowest-performing employees are presumably smaller. Therefore, we expect heterogeneous feedback effects only for co-worker-related information and propose the following:

Hypothesis 3. *The effect of real-time feedback containing "social" and "personal and social" performance averages is greater for employees at intermediate levels of performance and less for employees at the extreme ends.*

1.3 Methodology

1.3.1 Company setting

Our project partner is a railway catering enterprise in Switzerland. The largest company unit includes the service of meals, snacks, and drinks on Swiss trains by so-called stewards. By February 2016, the company employed 199 minibar stewards who sell drinks and snacks from a mobile vending cart and 314 restaurant stewards who serve customers in the train restaurants. The target group of our experiment is the minibar stewards. In contrast to the service personnel in the train restaurants, the minibar stewards are salespeople with a strong and direct influence on sales performance. They manage demand, for example, through their walking speed, friendliness, verbal promotion, and

cross-selling efforts.[2] The motivation and effort of the minibar stewards also plays a crucial role in customer satisfaction and the company's reputation in general.

Employee motivation is one of the company's major challenges. The job of the minibar stewards is not highly regarded, rather isolated, and repetitive. Due to the weight of the vending cart, the work is also physically demanding, which explains why 98% of the minibar stewards are male. Another management challenge is the lack of control mechanisms. Because the minibar stewards usually start, execute, and terminate their services alone, there is hardly any interaction with superiors or co-workers.

To manage employee motivation, the company currently applies an incentive scheme consisting of a fixed wage and a monetary reward for above-average sales performance. This system provides the prospect of significant bonus payments that, according to the company, account for up to 20% of a steward's monthly income. As revenues greatly depend on train routes and service times, the incentive scheme compares a steward's sales revenue to the average revenue of his co-workers on the same work shift. Before proceeding with the incentive scheme, Table 1.1 and Table 1.2 provide a brief outline of the main shift characteristics of our sample during the study period.

A shift starts and ends at a certain time at a certain destination (usually the steward's official place of employment) and covers a specific train route. During our study period, the company operated 104 different minibar shifts, starting at one of eight major Swiss train stations (see Table 1.1).

TABLE 1.1: Shifts and services per region

City of start	N (shifts)	N (services)	Percent (services)
Basel	12	659	10.66
Bern	14	1,215	19.66
Brig	5	565	9.14
Chur	4	206	3.33
Genf	9	738	11.94
Luzern	4	421	6.81
St. Gallen	3	206	3.33
Zürich	53	2,170	35.11
Total	104	6,180	100

Most shifts are performed on a daily basis. These daily assignments are referred to as a service, that is, a shift performed by a certain steward on a certain date. As set out in

[2]This was not only stated in various interviews with the partner company but is also reflected in the data. The variance partition coefficient, which compares the between-employee revenue variation to the overall revenue variation in the data, is 21% for the minibar stewards and 11% for the restaurant stewards.

Table 1.1, our dataset contains 6,180 minibar services that were performed on one of the 104 minibar shifts.

The shifts last an average of 9.1 service hours, of which 7.02 hours are effective working time (see Table 1.2). The stewards work on various shifts in accordance with the monthly deployment plan. Considering the study period, minibar stewards worked on average on 9.8 different minibar shifts (4.7 per month) and 5.6 times on the same shift (2.4 per month). The services are assigned by a separate planning department based on the stewards' place of employment and availability. According to the company, there is a tendency to assign well-performing stewards to busy shifts rather than poorly-performing stewards. The employees can state their shift preferences but have no direct influence on the service allocation.

TABLE 1.2: Shift characteristics

N=104	Mean	Min	Max
Shift duration	9.10	4.72	14.22
Work time	7.02	3.03	11.53
Break time	2.08	0.10	5.37
Different shifts per steward	9.80	1.00	32.00
Different shifts per steward/month	4.66	1.00	12.00
Same shift per steward	5.64	1.00	28.00
Same shift per steward/month	2.38	1.00	12.75

At the end of every month, a steward's personal average revenue of all his services on a certain shift is compared to the total average revenue of all stewards who have worked on the same shift. The weighted mean of these within-shift comparisons defines the steward's total performance in that month (mean deviation to the average shift revenues in %). With this approach, the company aims for a fair comparison of employees' productivity.[3] The stewards do not get to know their co-workers on a certain shift (i.e., their competitors) from the deployment plan and cannot strategically influence the shift assignments. Based on the monthly performance evaluation, the bonus pool is distributed as illustrated in Figure 1.1.[4]

Stewards receive a proportional bonus payment for above-average performance but no reward for below-average performance. This approach is similar to the proportional-prize

[3]To reduce the impact of extraordinary events or happenings, the performance measure is only calculated for stewards who have worked on 10 or more services per month.

[4]The volume of the bonus pool is confidential. It varies on average by 1%, depending on the overall sales per month.

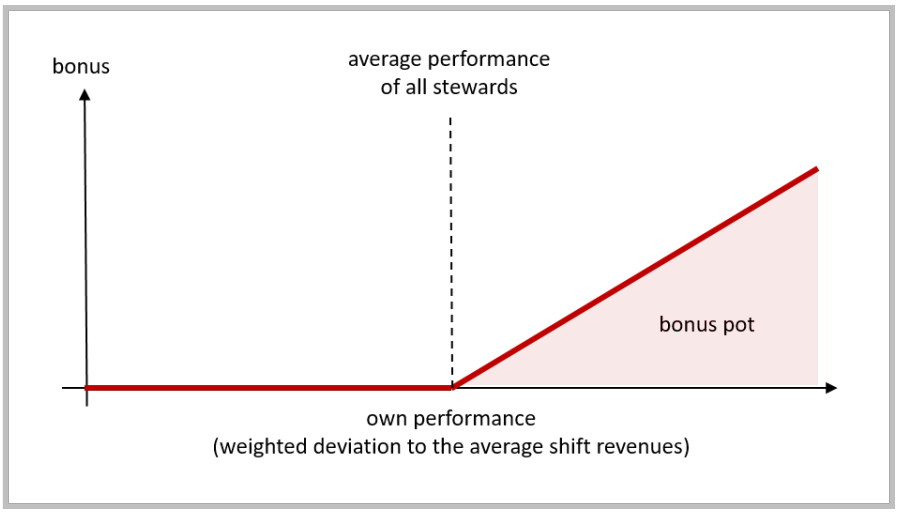


FIGURE 1.1: Incentive scheme

contest introduced by Cason et al. (2010), where the prize is distributed in proportion to the participants' achievement.[5]

At the end of every month, stewards are informed about their overall performance evaluation and the corresponding bonus payment on their salary statement. Apart from this performance summary, stewards hitherto received no regular feedback from superiors or any kind of revenue benchmarks. Our study was designed to exploit the motivational potential of ongoing, comparative performance feedback that is consistent with the incentive scheme.

1.3.2 Experimental design

We used a between-subject design consisting of three treatment groups and one control group. All treatment groups received regular feedback about the recent revenue averages of their current shift. This information was calculated in real time and appeared on the electronic checkout display of the vending cart. In the "personal info" treatment, stewards were informed about their own recent average, that is, the mean revenue of all services performed on the present shift during the last 30 days.[6] In the "social info" treatment, the message contained the recent average revenue of all stewards who worked

[5]The performance differences between stewards are quite large. During the 14-month pre-study period, the variation of the performance measure across minibar stewards goes from -22% up to 23%, with a standard deviation of 11.7 percentage points (see Figure 1.4, Appendix A). Performance variation per employee over time, however, is lower. The average standard deviation of a steward's performance over months lies at 7.9 percentage points. As expected, we observe far more performance variation between the minibar stewards than between restaurant stewards.

[6]If a steward did not work on the same shift in the recent past, the message still appeared but with an empty space. These occurrences were not considered in our analyses.

on the same shift during the last 30 days. Performance information of the "personal and social info" condition included both the shift-specific average of all workers, as well as the steward's personal average revenue on the same shift during the past 30 days. To rule out a behavioral change due to the messages per se, the control group received a general thank you message. Figure 1.2 shows an example of the messages received by the "personal and social info" treatment group and the control group. Both are translated in English.

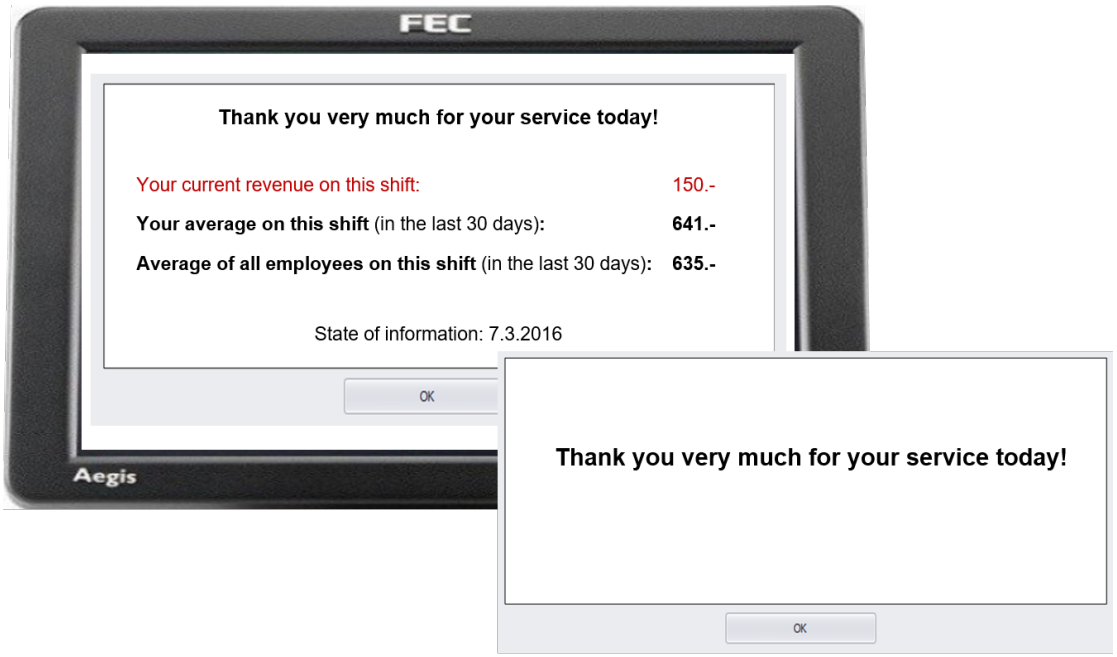


FIGURE 1.2: Translated message examples

Recall that the feedback in the "personal and social info" treatment is similar but more timely and more specific to what stewards receive in their monthly bonus accounting. Therefore, stewards cannot clearly infer monetary rewards from the feedback messages in either of the treatment groups. In contrast to the incentive scheme, the performance information was also dynamic and always referred to the last 30 days instead of comparing performance within the same month. We thereby ensured that the information is up to date, while keeping the "informational value" of the feedback message constant over time. Within-month feedback, in contrast, would have generated many empty or unreliable messages at the beginning of the month when the number of services performed on a certain shift is still small.

The information sent to the three treatment groups also contained a steward's current sales revenue that he hitherto generated on his service. Contrary to the revenue averages, stewards can access this information on the electronic tills at any time. Furthermore, the generated revenue automatically appears when stewards do the daily accounting at the end of their service. The feedback provided in our "personal info" treatment is therefore

less novel than the messages of the "social info" and "personal and social info" groups. As in most work settings, employees could theoretically track their personal performance, for example, by writing down the revenue after every service in our case.

The messages were programmed by an external IT company that also maintains the electronic till system of the project partner. The average personal and social sales revenues per shift were automatically calculated in real time when the stewards logged onto the till at the beginning of their service. Respective performance information appeared on the checkout display at three different times per day: at the beginning of the service (login), at the end of the service (logoff), and once at a random time during work. With this during-service feedback, we aimed to additionally explore the *immediate* performance effect of real-time feedback over the subsequent working hours. The corresponding analyses are provided in Appendix D. The thank you message for the control group only appeared once, at the beginning of the service. To ensure that stewards read the message, they had to click the "OK" button before they could proceed with another till transaction. Furthermore, language was adapted automatically, depending on the steward's reference language (German, French, or Italian).

Our intervention ran from March 1 to June 30, 2016. All 199 minibar stewards that were employed by February 1, 2016 were randomly assigned to one of the four experimental conditions. By stratification, we ensured a balanced distribution of the stewards' prior sales performance that may interact with our intervention.[7] During the study period, real-time feedback or thank you messages were provided on all services, except for extra or charter shifts. We also excluded foreign train connections, operated by TGV Lyria and SNCF Voyages Italia from the study, as these shifts have different service processes.

Importantly, stewards did not know that they were taking part in an experiment. Prior to launch, the participants were only informed that the head office was going to use tills more frequently as a communication channel. This information was also sent via the electronic cash desk. The eight sales managers (direct superiors of the minibar stewards) received a general e-mail from the human resource department informing them about the attempt to provide additional revenue information to stewards. It was also explained that this revenue information could vary during the initial test period of the project.

[7] As we had no other sales or bonus data available at that time, we used the bonus calculations of November and December 2015 as prior performance measures for stratification.

1.3.3 Field data and sample characteristics

Our dataset consists of all minibar services performed by the minibar stewards between January 1, 2015 and June 30, 2016. We refer to the time before the intervention, from January 1, 2015 until February 28, 2016 as pre-study period. The time from March 1 to June 30, 2016 is referred to as study period. In addition to individual sales data, we obtained confidential data on the daily passenger numbers per train from Swiss Federal Railways. This data was used to calculate the number of passengers per minibar service. Because the number of passengers on the trains is an important control variable in our analyses, we excluded services for which passenger data was incomplete.[8] We further omitted services that were affected by a train failure or that did not report any revenue, for example, due to a malfunction of the cash desk. During the study period, we also excluded those observations of the treatment groups where the performance information was incomplete or missing, for example, because the steward did not work on the same shift during the last 30 days. Using these specifications, we had to exclude 28.5% of the minibar services (and two stewards) during the study period. Incomplete performance information and missing passenger data accounted for most part of these cases.[9]

Our final data set contains 33,064 minibar service observations, 6,180 in the study period and 26,884 in the pre-study period. The service observations of the study period were performed by 164 minibar stewards, whereas 172 stewards were active during the whole observation period (January 1, 2015 to June 30, 2016). Table 1.3 provides an outline of the number of observations and stewards across the experimental conditions. The lower part of the table shows the main sample characteristics of the stewards and service observations during the pre-study period. Service-related variables report the means per service, whereas steward-related variables show the average values across stewards.

Most stewards are long-term employees with an average tenure of seven years. The average workload of the stewards before the intervention was 11.9 services per month (without services that were excluded from our data set). Furthermore, stewards worked on 5.4 different shifts and performed, on average, 2.5 services on the same shift per

[8] A minibar service covered between one and eight different trains. Passenger data was considered incomplete if there were one or more trains involved in a service for which passenger numbers were not recorded.

[9] Performance information was particularly incomplete in the "personal info" and "personal and social info" treatments. This is because the stewards did not necessarily work on the same shift during the last 30 days before the message release, leading to a missing personal average. Therefore, 30% (34%) of the study period observations in the "personal info" ("personal and social info") treatment had to be excluded. In the "social info" treatment, this figure was only 2%, which explains the higher number of services in this group.

TABLE 1.3: Sample characteristics

	Personal info	Social info	Personal + Social	Control	Sample
N (stewards) study period	39	42	41	42	164
N (stewards) overall	41	44	44	43	172
N (services) study period	1,242	1,902	1,291	1,745	6,180
N (services) overall	7,507	8,994	8,377	8,186	33,064
Steward characteristics (pre-study):					
Tenure (years)	6.92	5.51	7.05	8.54	7.00
	(5.75)	(4.87)	(5.15)	(7.40)	(5.92)
Workload (ø services per month)	11.68	12.47	12.00	11.56	11.94
	(3.95)	(4.37)	(4.04)	(4.93)	(4.32)
No. different shifts per month	5.60	5.32	5.54	5.11	5.39
	(2.12)	(2.34)	(2.26)	(2.22)	(2.23)
No. same shift per month	2.34	2.64	2.49	2.62	2.53
	(1.13)	(1.29)	(1.30)	(1.85)	(1.41)
Service characteristics (pre-study):					
Log revenue per hour (CHF)	3.97	3.93	3.97	3.96	3.96
	(0.41)	(0.41)	(0.41)	(0.39)	(0.4)
Log items sold per hour	2.46	2.42	2.47	2.45	2.45
	(0.42)	(0.43)	(0.43)	(0.41)	(0.42)
Log customers per hour	1.98	1.95	1.98	1.98	1.97
	(0.48)	(0.5)	(0.5)	(0.5)	(0.5)
Items sold per customer	1.75	1.78	1.78	1.82	1.79
	(2.02)	(2.85)	(2.79)	(4.09)	(3.04)
Worktime (hours)	6.47	6.58	6.53	6.49	6.52
	(1.92)	(2.14)	(1.96)	(1.77)	(1.96)
Break (hours)	1.90	1.67	1.84	1.75	1.79
	(1.3)	(1.17)	(1.24)	(1.21)	(1.23)
Train occupancy (%)	37.44	38.33	37.78	37.34	37.74
	(10.8)	(11.45)	(10.84)	(10.44)	(10.91)
Share 1st class passengers	18.60	18.51	18.60	18.45	18.54
	(5.59)	(5.77)	(5.66)	(5.47)	(5.63)

Notes: The table reports the descriptive statistics for each treatment group individually and for the full sample. Steward characteristics show the average values across stewards, whereas service characteristics show the average values across services (i.e., a shift performed by a certain steward on a certain date). All means refer to the pre-study period. Standard deviations are shown in parentheses.

month. The service-related variables further show the mean sample characteristics per service.

Our main outcome variable, sales performance, is defined as the logarithmized revenue per hour on each service (in Swiss francs, CHF). As shown in Figures 1.5 and 1.6 of Appendix A, this variable follows a normal distribution with a few downward outliers. We did not use an aggregated performance measure at the steward level as a dependent variable for several reasons. First, there are major concerns with aggregating hierarchical data structures. The loss of variance information at any level can lead to severely incomplete

or even misleading knowledge (Bullen et al. 1997, Subramanian et al. 2009). This risk is particularly high in our case, where we observe high variation on the lower level, that is, the services. Second, analyzing revenue averages on the steward level makes inference highly volatile. The results of such an analysis strongly depend on the exact specification of the performance measure. Third, taking a steward-related outcome measure that controls for shift differences (e.g., the bonus calculation) entails an endogeneity problem; a steward's performance in this case depends on the performance of the other employees working on the same shift. According to our hypotheses, the other stewards' performance in turn depends on their assignment to the treatment groups. We therefore conduct our analyses on the level of services, using each service as a single observation.

1.4 Results

1.4.1 Real-time feedback and sales performance

To investigate the effect of real-time feedback on sales performance, we follow the approaches of Gneezy and List (2006) and Friebel et al. (2017) by using a random intercept model with random effects for stewards (see Cameron and Trivedi 2010, pp. 232-256).[10] In our regression model, the logarithmized revenue per hour of service i, j, t (i.e., the shift j performed by steward i on date t) is defined as:

$$\begin{aligned} Ln(revhour)_{i,j,t} = \beta_0 + \beta_1 Group_i + \beta Stew'_i + \beta Shift'_j + \beta Date'_t \\ + \beta Service'_{i,j,t} + \upsilon_i + \epsilon_{i,j,t}. \end{aligned} \tag{1.1}$$

The variable $Group_i$ is a categorical variable with four levels (three treatments and one control group), identifying the experimental condition of steward i. Besides this main variable of interest, we include multiple control variables referring to the steward-, shift-, date- and service-specific characteristics of our service observations. $Stew'$ is a vector containing steward-specific controls. These are tenure, workload (average number or services per month), and employment status (temporary or permanent). To control for a steward's general ability, we also integrate an indicator for the average sales performance of steward i before the intervention. This measure is computed in the same manner as

[10] Recall that a service is defined as a shift performed by a certain steward on a certain date. Therefore, our service observations are nested within shifts and within stewards, that is, a cross-classified data structure with two levels. In our analyses, we consider the services as the first level of analysis and the stewards as level two. Besides steward characteristics, we include shift- and date-related control variables that refer to the service level.

the monthly bonus calculation of the partner company (i.e., the mean deviation between the personal and overall revenue averages per shift, see Section 1.3.1).[11]

The vector $Date'$ includes time-dependent covariates that presumably influence consumption on the trains. These are dummy variables for the months and an indicator for weekends or holidays versus business days. $Shift'$ is a vector containing shift-related controls, that is, information associated with a the shift that steward i performs on date t. These controls include the type of the shift (i.e, whether there is a restaurant or a bistro on the train), the city of shift start and shift duration (work time). We also created a variable indicating to what extent the shift covers common eating times, meaning breakfast, lunch, and dinner times in % of total work time. In addition to these shift-related variables, we control for other service-related characteristics $Serv'$ that are shift- *and* date-specific. These variables include the average train occupancy of the $\text{service}_{\text{i,j,t}}$ and the average share of 1st class passengers. We found that these variables together with the steward- and shift-related characteristics explain 86% of the between-steward variance in sales revenues.[12] Occupancy shows the percentage share of occupied train seats (mean over all trains that are involved in the service) and was computed using confidential data on the daily passenger numbers of Swiss Federal Railways.[13] The reason we control for several date-, shift-, and service-related variables in addition to the occupancy rate is that they presumably affect consumption patterns beyond the mere amount of passengers. For example, it is likely that passengers consume more during weekends or that passenger types and spending behavior vary with respect to the city of shift start. Finally, we take into account whether a second steward was working on a particular service (which was the case for only 11 observations during the study period) and whether the service was affected by a major event near the service route. The last two terms of Model 1.1 indicate random steward-specific deviations from the average (υ_i) and the random error ($\epsilon_{i,j,t}$).

[11] To calculate a steward's prior performance, we took the weighted average of the monthly bonus calculations over the pre-study period.

[12] Including 104 single-shift dummies, instead of the shift-related variables and passenger numbers, does not improve the fit of our model for the between-steward differences (R^2_{btw} with dummies=0.857, R^2_{btw} without dummies=0.858). They rather absorb any individual steward effects, leading to a residual between-steward variance of σ_{u}=0. We explain this by the relatively low amount of different shifts per steward during the study period. Shift and steward performance may also be interdependent, if well-performing stewards are rather assigned to busy shifts (see Section 1.3.1). We therefore adhere to more precise and more efficient occupancy measure to control for the sales potential of the service. Also see Breheny (2017) for an overview of overfitting problems.

[13] We did not use absolute passenger numbers but occupancy rates, as we want to model a non-linear relationship between the share of occupied seats and sales revenues. We presume lower sales in very crowded trains. The expected non-linear relationship between the number of passengers and sales performance is also the reason why we did not use the revenue per passenger as an outcome measure in our analyses.

Table 1.4 provides the estimates of Model 1.1 during the study period. For parsimony, we excluded variables with no significant effects, which were stewards' tenure, employment status, and workload. Including these controls has a negligible influence on the results. Steward-, shift-, date-, and service-related control variables were sequentially added in Specifications (2), (3), (4), and (5). Cluster-robust standard errors are shown in parentheses.

TABLE 1.4: Random effects regression: Log revenue per hour

	(1)	(2)	(3)	(4)	(5)
personal info	-0.0055	0.0134	0.0215	0.0205	0.0170
	(0.0401)	(0.0321)	(0.0163)	(0.0161)	(0.0170)
social info	-0.0200	0.0068	0.0292**	0.0297**	0.0329***
	(0.0362)	(0.0285)	(0.0143)	(0.0139)	(0.0127)
personal + social	0.0160	0.0277	0.0328*	0.0326*	0.0393**
	(0.0410)	(0.0350)	(0.0197)	(0.0191)	(0.0155)
performance before (in %)		0.0110***	0.0103***	0.0103***	0.0101***
		(0.0012)	(0.0006)	(0.0006)	(0.0006)
worktime (in h)			-0.0305***	-0.0307***	-0.0137***
			(0.0034)	(0.0034)	(0.0032)
eating times (in %)			0.0046***	0.0046***	0.0076***
			(0.0008)	(0.0008)	(0.0007)
weekend/holiday				-0.0205*	0.1662***
				(0.0114)	(0.0145)
occupancy (in %)					0.0245***
					(0.0028)
occupancy^2					-0.0001***
					(0.0000)
1st class pass (in %)					0.0095***
					(0.0012)
no. stewards working					-0.7318***
					(0.1892)
event					0.1399***
					(0.0304)
shift type effects	No	No	Yes	Yes	Yes
city of shift start	No	No	Yes	Yes	Yes
month effects	No	No	No	Yes	Yes
sd (stewards)	0.172	0.131	0.055	0.047	0.021
sd (residual)	0.333	0.333	0.319	0.318	0.291
R^2 overall	0.001	0.074	0.247	0.254	0.376
Observations	6,180	6,149	6,149	6,149	6,149
N stewards	164	162	162	162	162

Notes: Generalised least squares (GLS) regression of the logarithmized revenue per hour (per service) with random effects for stewards. Robust standard errors clustered on the individual level are in parentheses. *personal info*, *social info*, and *personal + social* are dummy indicators for the experimental treatments, whereas the control group is the reference category. By adding shift-, date-, and service-related control variables, Specifications (3), (4), and (5) also include fixed effects for the shift type, the city of shift start, and month. See the discussion of Model 1.1 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As indicated by the low R^2 in Specification (1), we are trying to estimate a rather weak signal in the presence of a lot of noise. Without any control variables, our estimates therefore do not reveal any treatment effect. The treatment coefficients become positive when controlling for the stewards' prior performance in Specification (2) and significant for the "social" and "personal and social info" treatment as soon as we control for shift-related characteristics in Specification (3). The shift-related control variables also significantly improve the fit of the model. Adding date- and service-related characteristics in Specifications (4) and (5) further increases the effect sizes. The results in Specification (5) show a significant increase in the revenue per hour of 3.3% (3.9%) for services performed by stewards in the "social info" ("personal and social info") treatment group compared to the control group. The effect of the "personal info" condition is also positive but not significant. All the other coefficients point in the expected directions. The differences between the treatment groups are not significant (p=0.32 "personal info" vs. "social info", p=0.209 "personal info" vs. "social info", p=0.649 "social info" vs. "personal and social info", Wald test).

Table 1.5 provides the estimates of Specification (5) for additional outcome measures, such as the number of items sold and the number of different transactions (i.e., customers served) per hour. All variables are logarithmized.

As shown in Specification (2) of Table 1.5, the treatment effects become even more evident when considering the number of products sold. The number of items sold per hour is up to 4.7% (4%) higher in the "personal and social info" ("social info") condition than in the control group. The coefficients of the "personal and social info" treatment in the last two columns indicate that this effect can mainly be attributed to a higher number of customers rather than to enhanced cross-selling activities with additional products sold per customer. The fact that performance differences are particularly driven by the number of customers is also reflected in our pre-intervention data. Top performers do not sell more products per customer than poor-performing stewards but reach more buyers. However, this effect is less clear for the "social info" treatment, indicating that stewards may have individual sales strategies for increasing their revenue.

The results above are stable when conducting various robustness checks. Table 1.7 in Appendix B reveals very similar results for an ordinary least squares (OLS) regression with pooled service data and cluster-robust standard errors at the steward level.[14] Following the approaches of Friebel et al. (2017) and Kallbekken and Sælen (2013) with a comparable data structure, we also perform a difference-in-difference analysis. The

[14]Clustered standard errors are used to control for heteroskedasticity and correlation of errors within stewards across services (see Colin Cameron and Miller 2015).

TABLE 1.5: Random effects regression: Log revenue, items, and customers

	(1) log revenue per hour	(2) log items per hour	(3) log customers per hour	(4) log items per customer
personal info	0.0170	0.0203	0.0044	0.0156
	(0.0170)	(0.0173)	(0.0194)	(0.0110)
social info	0.0329***	0.0397***	0.0221	0.0174*
	(0.0127)	(0.0138)	(0.0162)	(0.0099)
personal + social	0.0393**	0.0474***	0.0411**	0.0055
	(0.0155)	(0.0169)	(0.0191)	(0.0093)
performance before (in %)	0.0101***	0.0103***	0.0103***	-0.0001
	(0.0006)	(0.0006)	(0.0007)	(0.0003)
worktime (in h)	-0.0137***	-0.0077**	-0.0276***	0.0200***
	(0.0032)	(0.0036)	(0.0040)	(0.0022)
eating times (in %)	0.0076***	0.0124***	0.0099***	0.0025***
	(0.0007)	(0.0008)	(0.0008)	(0.0004)
weekend/holiday	0.1662***	0.1714***	0.1096***	0.0619***
	(0.0145)	(0.0149)	(0.0155)	(0.0087)
occupancy (in %)	0.0245***	0.0245***	0.0211***	0.0034***
	(0.0028)	(0.0030)	(0.0030)	(0.0013)
occupancy^2	-0.0001***	-0.0001***	-0.0001***	-0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
1st class pass (in %)	0.0095***	0.0109***	0.0128***	-0.0019**
	(0.0012)	(0.0013)	(0.0014)	(0.0008)
no. stewards working	-0.7318***	-0.7330***	-0.6915***	-0.0418
	(0.1892)	(0.1690)	(0.1734)	(0.0492)
event	0.1399***	0.1263***	0.0994***	0.0269**
	(0.0304)	(0.0296)	(0.0286)	(0.0120)
shift type effects	Yes	Yes	Yes	Yes
city of shift start	Yes	Yes	Yes	Yes
month effects	Yes	Yes	Yes	Yes
sd (stewards)	0.021	0.033	0.039	0.019
sd (residual)	0.291	0.297	0.326	0.174
R^2 overall	0.376	0.390	0.399	0.183
Observations	6,149	6,137	6,137	6,137
N stewards	162	162	162	162

Notes: GLS regression of the logarithmized revenue per hour (per service) with random effects for stewards. Robust standard errors clustered on the individual level are in parentheses. *personal info*, *social info*, and *personal+social* are dummy indicators for the experimental treatments, whereas the control group is the reference category. All specifications also include fixed effects for the shift type, the city of shift start, and month. See the discussion of Model 1.1 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

estimates in Table 1.8 of Appendix B demonstrate that: 1) we obtain similar results when comparing the pre- and during-study periods with significant treatment effects for the "social info" and the "personal and social info" groups; 2) the effects also persist when using fixed instead of random effects at the steward level; and 3) the results of the difference-in-difference model are also robust towards modifications in the control variables. We further tested whether the productivity increase could be attributed to a short-term enhancement of motivation when feedback was launched at the beginning of the study period. As shown in Table 1.9 of Appendix C, we find no significant interaction effect between the "social info" or "personal and social info" treatment groups and the days after the study start. The performance increases seem to persist over time. Our results are also unlikely to be a consequence of changes in the workforce. The analysis only includes employees who were recruited at least one month before the start of the experiment and our data does not reveal an increased drop-out rate for poor-performing employees in the treatment groups during the study period.[15]

Overall, we observe a quantitatively large and statistically significant effect of real-time feedback that contains recent, co-worker-related performance averages. Giving employees the opportunity to regularly compare themselves to their colleagues particularly increases the number of products sold. The related revenue increase that we observe in our experiment is comparable to the sales performance effects documented by Friebel et al. (2017) and Delfgaauw et al. (2013) using monetary incentives.

In contrast to messages containing social performance information, personal information alone had a positive but not significant effect on sales productivity. Our Hypothesis 1 is therefore only partially confirmed. However, the results are in line with Hypothesis 2, stating that the expected effects are lowest for the "personal info" group and strongest for the "personal and social info" group. We suggest that this result is driven by the fact that personal average performance levels are less novel and not bonus-relevant (see Sections 1.3.1 and 1.3.2).

Yet, the performance effects in our study are presumably not solely caused by the prospect of monetary rewards. This idea is supported by the fact that the incentive scheme is highly complex and stewards cannot directly infer a financial bonus from the feedback messages. Furthermore, we do not observe a more powerful impact of our treatments toward the end of the month when social performance information (showing the last 30-day averages) is closest to the performance benchmark used for the bonus calculation.

[15] We have no clean data on withdrawals for the study participants. However, the service observations show that only three employees of the treatment groups did not work during the last or the last two months of the study period. Therefore, there are at maximum three employees that possibly left the company during the intervention, and these are not necessarily poor performers.

As shown in Table 1.10 of Appendix C, the regression coefficients for the treatment and day-of-month interactions are very low and not significant. We even observe a slight performance decrease when comparing the average revenues per hour in the middle and at the end of the month (middle and last 10 days) to those at the beginning (first 10 days) of the month within the treatment groups. We therefore suggest that psychological factors that may arise from relative comparisons, such as self-satisfaction and self-efficacy (Bandura and Cervone 1983, Bandura 1988) or conformity effects (Bernheim 1994) are also important for explaining our results.

1.4.2 Real-time feedback effects and ability

As set out in Section 1.2.3, we presume different reactions to the feedback messages, depending on a steward's general level of ability. To test this hypothesis, we split the minibar stewards into four performance quartiles: the worst 25%, the worse 25%, the better 25%, and the best 25%. These quartiles are based on the stewards' prior sales performance in the pre-study period (see Section 1.4).[16] With reference to Model 1.1, we estimate the following interaction effects:

$$Log(revhour)_{i,j,t} = \beta_0 + \beta_1(Group_i * Quartile_i) + \beta Stew'_i + \beta Shift'_j + \beta Date'_t + \beta Service'_{i,j,t} + \upsilon_i + \epsilon_{i,j,t}. \quad (1.2)$$

$Log(revhour)_{i,j,t}$ is the logarithmized revenue per hour on service i, j, t, that is, the hourly revenue achieved by steward i on shift j on date t. $Group_i * Quartile_{\text{i}}$ are the interaction terms for each treatment group with each performance quartile. All control variables are equal to Model 1.1 (see Specification 5 in Table 1.4). υ_i indicates the random effects for stewards, and $\epsilon_{i,j,t}$ is the idiosyncratic error term which is clustered at the steward level. Table 1.6 provides the estimates of Model 1.2.

In line with Hypothesis 3, the interaction coefficients are particularly high and significant for the performance quartiles around the median. For the worse 25% of the stewards, the "personal and social info" treatment, for example, leads to an increase in revenue per hour of up to 15% compared to the worst 25% in the control group (reference category). As the first three rows of the regression output reveal, the treatment effects for the poorest performers tend to be negative. The treatment coefficients for the best 25% are positive but except for the "personal and social info" condition not significant. According

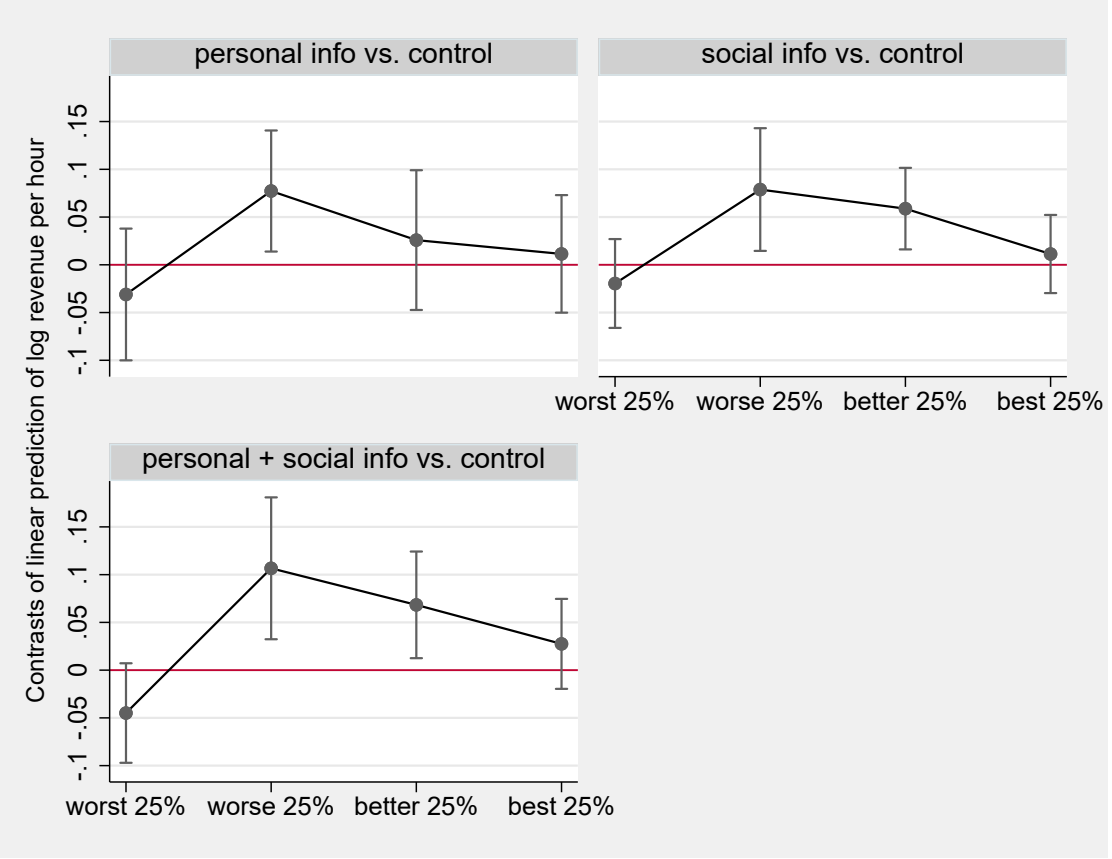
[16] Using this performance measure instead of a more recent or dynamic indicator allows us to uniquely assign each employee to one of the four performance groups and avoids endogenous interactions with our intervention.

TABLE 1.6: Random effects regression: Treatment–performance interactions

	(1) log revenue per hour
personal info	-0.0311
	(0.0352)
social info	-0.0196
	(0.0237)
personal + social	-0.0449*
	(0.0266)
worse 25% x personal info	0.1083**
	(0.0463)
worse 25% x social info	0.0984**
	(0.0407)
worse 25% x personal + social	0.1516***
	(0.0467)
better 25% x personal info	0.0569
	(0.0516)
better 25% x social info	0.0784**
	(0.0331)
better 25% x personal + social	0.1133***
	(0.0398)
best 25% x personal info	0.0425
	(0.0462)
best 25% x social info	0.0309
	(0.0313)
best 25% x personal + social	0.0724**
	(0.0348)
steward controls	Yes
shift controls	Yes
date controls	Yes
service controls	Yes
sd (stewards)	0.011
sd (residual)	0.291
R^2 overall	0.379
Observations	6,149
N stewards	162

Notes: The table displays the estimates of the logarithmized revenue per hour (per service), using a GLS regression with random effects for stewards. Robust standard errors clustered on the individual level are in parentheses. *personal info*, *social info*, and *personal+social* are dummy indicators for the experimental treatments, whereas the control group is the reference category. All steward-, shift-, date-, and service-related control variables are included. See the discussions of Models 1.1 and 1.2 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

to the Wald tests, the feedback effects particularly differ between the worst and the worse 25% of the stewards and between the worst and the better 25%. Within the performance quartiles, however, the three treatment groups have no significant different effect on performance. To illustrate the effect sizes, Figure 1.3 shows the differences in the predicted margins for each treatment group and each performance quartile.



Notes: For each performance quartile (based on a steward's performance in the pre-study period), the graphs show the estimated marginal effect (Model 1.2) of the treatment groups compared to the control group. The error bars indicate the 95% confidence intervals.

FIGURE 1.3: Contrasts of predictive margins of Model 1.2

As indicated in Table 1.6, the treatment effects on the logarithmized revenue per hour are particularly strong for those stewards who usually perform just below average. Here, we observe a productivity increase of up to 10% ("personal and social info") compared to the control group (p=0.005, Wald test). However, the post-estimation tests confirm that all types of real-time feedback have no significant effect on revenues for the best- and poorest-performing stewards.[17] The results are similar when using the difference-in-difference approach with fixed effects for stewards as discussed in Section 1.4.1 (see Figure 1.7 in Appendix B).

[17] Only the worst 25% in the "personal and social info" group show a negative reaction at the 10% significance level.

While these findings basically meet our expectations as stated in Hypothesis 3, some outputs stimulate further discussion. In contrast to previous studies on the dynamic incentive effect, stewards at the extremes of the performance distribution are not negatively affected by our feedback intervention. We believe that this can be attributed to two characteristics of our design. First, relative incentives in our experiment rather resemble a multi-stage proportional prize contest than a tournament with one or a few winners (e.g., Hannan et al. 2008, Delfgaauw et al. 2014). This feature presumably mitigates a negative effect among the top and lowest performers. Second, the indirect link between performance feedback and monetary rewards in our setting may be a supportive factor in the sense that participants cannot directly infer monetary consequences from behavioral changes. We suppose that this has a similar positive effect as partial disclosure policies or vague feedback. Both were proposed to maintain motivation for top and low performers in earlier studies (Hannan et al. 2008, Goltsman and Mukherjee 2011).

We further observe a similar pattern of heterogeneous feedback effects in all treatment groups. While this seems surprising, it supports our previous point that the performance improvements in our study are probably not only driven by rational considerations (i.e., potential bonus payments) but by behavioral factors as well. The "personal info" treatment, for example, does not offer any reward-related information but still has a significant positive impact on the worse 25% of the stewards. The frequent tracking of revenues and the enhanced concern about performance seems to motivate just below-average performers to realize their potential. Very poorly performing employees, however, are rather discouraged by receiving any type of comparative performance information. According to the company, direct superiors already exert considerable pressure on stewards with continuing low sales figures. We therefore presume that these employees may hardly improve with any kind of feedback. Similarly, none of the feedback messages lead to a revenue increase for the best 25% on the other side of the performance distribution. Although the monthly bonus is proportionally distributed and additional effort would therefore pay off, we do not observe any significant treatment effects in this performance quartile. In line with previous evidence (e.g., Eriksson et al. 2009), we explain this result with a ceiling effect, suggesting that the top 25% of the stewards have already been working close to their performance limit.

1.5 Discussion

Practitioners increasingly recognize the benefits of providing frequent and timely performance evaluations to employees (Duggan 2015). Yet, scientific evidence around the impact and optimal design of real-time feedback is surprisingly scarce. This study is one

of the first contributions in the field, confirming that real-time performance information can indeed lead to a significant productivity increase beyond what is achieved by traditional feedback. In the presence of a relative incentive scheme, our results show a lasting growth in sales revenues of up to 3.9% when employees are regularly informed about personal and co-worker-related performance averages of their current work task. This information is given in addition to an aggregated performance signal at the end of every month. Timely co-worker-related performance information alone leads to similar improvements. Providing real-time feedback only about personal performance standards, however, has no significant effect on sales productivity in our setting.

These results indicate that in competitive environments, productivity is influenced by timely and privately observed information about the performance of peers. This is in line with existing evidence around social comparative information and rank feedback, suggesting that giving people the opportunity to compare themselves to others can elicit considerable productivity gains (e.g., Szymanski and Harkins 1987, Blanes i Vidal and Nossol 2011, Kuhnen and Tymula 2012). Our findings also add to previous studies that propose peer monitoring as an effective incentive mechanism at work (Falk and Ichino 2006, Mas and Moretti 2009). Our results suggest that output can be similarly increased when co-worker-related performance information is frequently revealed in an individual work setting where feedback is private.

The productivity growth in our study can be traced to the fact that employees sell more products to a larger number of customers rather than selling more expensive items or intensified cross-selling. This is consistent with earlier work, suggesting that competitive incentives may induce individuals to work harder but not necessarily smarter (Casas-Arce and Martínez-Jerez 2009, Bracha and Fershtman 2013).

Our study also offers insights into how comparative performance information interacts with employees' general levels of performance. The productivity increases in our intervention are driven by workers in the middle of the performance distribution, especially by those who usually perform just below the median. Building upon the literature on dynamic incentive effects (e.g., Casas-Arce and Martínez-Jerez 2009, Bandiera et al. 2013, Delfgaauw et al. 2014) and self-confidence and self-efficacy theory (e.g., Bandura and Cervone 1983, Benabou and Tirole 2002), this finding confirms the non-linear relationship between a worker's performance–standard discrepancy and his or her subsequent effort. Organizations may therefore strategically use relative performance information, for example, by adapting the frequency of feedback or by using different reference groups, depending on an employee's general level of performance (see Kuhnen and Tymula 2012).

From a practical perspective, the monetary gains of timely co-worker-related performance information are quite substantial. In our study, an increase of 3.9% in revenue per hour equals approximately 34,000 CHF additional revenue per month. Interestingly, and important from a practical point of view, the positive impact of real-time feedback does not seem to fade over time. Assuming that the effect is persistent, the monthly benefits correspond to a revenue growth of more than 400,000 CHF per year. Furthermore, this increase in productivity comes at almost no cost. The one-off expenditures for our intervention were only 15,000 CHF for message programming. As the existing incentive scheme is based on relative performance, the company also does not face additional bonus expenses.

Our interpretation of the results is that the productivity improvements in our study were triggered by rational and psychological implications of the real-time feedback messages. We presume that the prospect of monetary rewards and concerns about relative performance per se supported the effects. The role of different incentive schemes and other behavioral factors related to our results needs to be explored in future research. Importantly, future studies should also investigate setting-related aspects that we could not consider in this experiment. Gender effects, for example, may have a significant influence on the outcome of timely performance information that allows social comparisons (Barankay 2011a, Delfgaauw et al. 2013). With the data at hand, we seek further insights on the impact of different performance–standard discrepancies during work and the immediate influence of benchmark achievements on performance. As firms increasingly adapt their feedback practices, these and other questions related to real-time feedback remain of great interest.

References

Akın, Z. and Karagözoğlu, E. (2017). The role of goals and feedback in incentivizing performance. *Managerial and Decision Economics*, 38(2):193–211.

Alavosius, M. P. and Sulzer-Azaroff, B. (1986). The effects of performance feedback on the safety of client lifting and transfer. *Journal of Applied Behavior Analysis*, 19(3):261–267.

Alvero, A. M., Bucklin, B. R., and Austin, J. (2001). An objective review of the effectiveness and essential characteristics of performance feedback in organizational settings (1985-1998). *Journal of Organizational Behavior Management*, 21(1):3–29.

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

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., and Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, 27(1):166–177.

Azmat, G. and Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy*, 25(1):77–110.

Bandiera, O., Barankay, I., and Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, 11(5):1079–1114.

Bandura, A. (1988). Self-regulation of motivation and action through goal systems. In Hamilton, V., Bower, G. H., and Frijda, N. H., editors, *Cognitive Perspectives on Emotion and Motivation*, pages 37–61. Springer Netherlands, Dordrecht.

Bandura, A. and Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45(5):1017–1028.

Bandura, A. and Jourden, F. J. (1991). Self-regulatory mechanisms governing the impact of social comparison on complex decision making. *Journal of Personality and Social Psychology*, 60(6):941–951.

Barankay, I. (2011a). Gender differences in productivity responses to performance rankings: Evidence from a randomized workplace experiment. Working Paper, Wharton School.

Barankay, I. (2011b). Rankings and social tournaments: Evidence from a crowd-sourcing experiment. Working Paper, University of Pennsylvania.

Bartel, A. P. (2004). Human resource management and organizational performance: Evidence from retail banking. *Industrial and Labor Relations Review*, 57(2):181.

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

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

Blader, S., Gartenberg, C. M., and Prat, A. (2015). The contingent effect of management practices. Columbia Business School Research Paper No. 15-48.

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

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

Breheny, P. (2017). *Advanced Regression*: *Model selection 1*. Retrieved from https://web.as.uky.edu/statistics/users/pbreheny/760/S11/notes/2-17.pdf.

Bullen, N., Jones, K., and Duncan, C. (1997). Modelling complexity: Analysing between-individual and between-place variation—a multilevel tutorial. *Environment and Planning A*, 29(4):585–609.

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics using Stata*. A Stata Press publication. Stata Press, College Station, Tex.

Cappelli, P. and Tavis, A. (2016). The performance management revolution. *Harvard Business Review*, 94(10):58–67.

Casas-Arce, P., Lourenço, S. M., and Martínez-Jerez, F. A. (2017). The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction. *Journal of Accounting Research*, 55(5):1051–1088.

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

Cason, T. N., Masters, W. A., and Sheremeta, R. M. (2010). Entry into winner-take-all and proportional-prize contests: An experimental study. *Journal of Public Economics*, 94(9-10):604–611.

Chhokar, J. S. and Wallin, J. A. (1984). A field study of the effect of feedback frequency on performance. *Journal of Applied Psychology*, 69(3):524–530.

Clark, A. E., Frijters, P., and Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the easterlin paradox and other puzzles. *Journal of Economic Literature*, 46(1):95–144.

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

Crowell, C. R., Anderson, D. C., Abel, D. M., and Sergio, J. P. (1988). Task clarification, performance feedback, and social praise: Procedures for improving the customer service of bank tellers. *Journal of Applied Behavior Analysis*, 21(1):65–71.

Delfgaauw, J., Dur, R., Non, A., and Verbeke, W. (2014). Dynamic incentive effects of relative performance pay: A field experiment. *Labour Economics*, 28:1–13.

Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326.

Deloitte (2015). *Global human capital trends 2015*: *Leading in the new world of work*. Deloitte University Press. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/de/Documents/human-capital/HCTrends%202015%20Report_TuesFeb24.pdf.

Duggan, K. (2015). Why the annual performance review is going extinct. *Fast Company*. Retrieved from https://www.fastcompany.com/3052135/why-the-annual-performance-review-is-going-extinct.

Earley, P. C., Northcraft, G. B., Lee, C., and Lituchy, T. R. (1990). Impact of process and outcome feedback on the relation of goal setting to task performance. *Academy of Management Journal*, 33(1):87–105.

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

Englmaier, F., Roider, A., and Sunde, U. (2017). The role of communication of performance schemes: Evidence from a field experiment. *Management Science*, 63(12):4061–4080.

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

Fajfar, P., Campitelli, G., and Labollita, M. (2012). Effects of immediacy of feedback on estimations and performance. *Australian Journal of Psychology*, 64(3):169–177.

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

Feather, N. T., editor (1982). *Expectations and Actions: Expectaney-Value Models in Psychology*. Lawrence Elbaum Associates, Hillsdale, New Jersey.

Fischer, C. (2008). Feedback on household electricity consumption: A tool for saving energy? *Energy Efficiency*, 1(1):79–104.

Friebel, G., Heinz, M., Krueger, M., and Zubanov, N. (2017). Team incentives and performance: Evidence from a retail chain. *American Economic Review*, 107(8):2168–2203.

Gneezy, U. and List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5):1365–1384.

Goltsman, M. and Mukherjee, A. (2011). Interim performance feedback in multistage tournaments: The optimality of partial disclosure. *Journal of Labor Economics*, 29(2):229–265.

Goomas, D. T. and Ludwig, T. D. (2007). Enhancing incentive programs with proximal goals and immediate feedback. *Journal of Organizational Behavior Management*, 27(1):33–68.

Goomas, D. T., Smith, S. M., and Ludwig, T. D. (2011). Business activity monitoring: Real-time group goals and feedback using an overhead scoreboard in a distribution center. *Journal of Organizational Behavior Management*, 31(3):196–209.

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

Heckhausen, H. (1977). Achievement motivation and its constructs: A cognitive model. *Motivation and Emotion*, 1(4):283–329.

Hedeker, D. and Gibbons, R. D., editors (2006). *Longitudinal Data Analysis*. John Wiley & Sons, Hoboken, NJ, USA.

Houde, S., Todd, A., Sudarshan, A., Flora, J. A., and Armel, K. C. (2013). Real-time feedback and electricity consumption: A field experiment assessing the potential for savings and persistence. *The Energy Journal*, 34(1):87–102.

Hundal, P. S. (1969). Knowledge of performance as an incentive in repetitive industrial work. *Journal of Applied Psychology*, 53(3):224–226.

Ilgen, D. R., Fisher, C. D., and Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4):349–371.

Jung, J. H., Schneider, C., and Valacich, J. (2010). Enhancing the motivational affordance of information systems: The effects of real-time performance feedback and goal setting in group collaboration environments. *Management Science*, 56(4):724–742.

Kallbekken, S. and Sælen, H. (2013). Nudging hotel guests to reduce food waste as a win–win environmental measure. *Economics Letters*, 119(3):325–327.

Kang, K., Oah, S., and Dickinson, A. M. (2005). The relative effects of different frequencies of feedback on work performance. *Journal of Organizational Behavior Management*, 23(4):21–53.

Kettle, K. L. and Häubl, G. (2010). Motivation by anticipation: expecting rapid feedback enhances performance. *Psychological Science*, 21(4):545–547.

Kim, J. S. and Hamner, W. C. (1976). Effect of performance feedback and goal setting on productivity and satisfaction in an organizational setting. *Journal of Applied Psychology*, 61(1):48–57.

Klein, W. M. (1997). Objective standards are not enough: Affective, self-evaluative, and behavioral responses to social comparison information. *Journal of Personality and Social Psychology*, 72(4):763–774.

Kluger, A. N. and DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284.

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

Lizzeri, A., Meyer, M. A., and Persico, N. (2002). The incentive effects of interim performance evaluations. CARESS Working Paper 02-09, University of Pennsylvania.

Locke, E. A., Shaw, K. N., Saari, L. M., and Latham, G. P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin*, 90(1):125–152.

Ludwig, S. and Luenser, G. (2008). Knowing the gap: Intermediate information in tournaments. Working Paper, University of Bonn.

Ludwig, T. D. and Goomas, D. T. (2009). Real-time performance monitoring, goal-setting, and feedback for forklift drivers in a distribution centre. *Journal of Occupational and Organizational Psychology*, 82(2):391–403.

Lurie, N. H. and Swaminathan, J. M. (2009). Is timely information always better? the effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes*, 108(2):315–329.

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

Mason, M. A. and Redmon, W. K. (2008). Effects of immediate versus delayed feedback on error detection accuracy in a quality control simulation. *Journal of Organizational Behavior Management*, 13(1):49–83.

McGregor, J. (2006). The struggle to measure performance. *Bloomberg Businessweek*. Retrieved from https://www.bloomberg.com/news/articles/2006-01-08/the-struggle-to-measure-performance.

Moore, D. A. and Klein, W. M. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes*, 107(1):60–74.

Nolan, T. V., Jarema, K. A., and Austin, J. (1999). An objective review of the journal of organizational behavior management. *Journal of Organizational Behavior Management*, 19(3):83–114.

Northcraft, G. B., Schmidt, A. M., and Ashford, S. J. (2011). Feedback and the rationing of time and effort among competing tasks. *The Journal of Applied Psychology*, 96(5):1076–1086.

PricewaterhouseCoopers [PwC] (2015). *The changing performance management paradigm: evolution or revolution?* Retrieved from https://www.pwc.nl/nl/assets/documents/pwc-performance-survey-2015.pdf.

Schultz, P. W. (1999). Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and Applied Social Psychology*, 21(1):25–36.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Schunk, D. H. and Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology*, 18(3):337–354.

Sharma, D. A., Chevidikunnan, M. F., Khan, F. R., and Gaowgzeh, R. A. (2016). Effectiveness of knowledge of result and knowledge of performance in the learning of a skilled motor activity by healthy young adults. *Journal of Physical Therapy Science*, 28(5):1482–1486.

So, Y., Lee, K., and Oah, S. (2013). Relative effects of daily feedback and weekly feedback on customer service behavior at a gas station. *Journal of Organizational Behavior Management*, 33(2):137–151.

Son, H. (2017). At JPMorgan, your performance review is now. and now. and now... *Bloomberg Businessweek*. Retrieved from https://www.bloomberg.com/news/articles/2017-03-09/at-jpmorgan-your-performance-review-is-now-and-now-and-now.

Strang, H. R., Lawrence, E. C., and Fowler, P. C. (1978). Effects of assigned goal level and knowledge of results on arithmetic computation: A laboratory study. *Journal of Applied Psychology*, 63(4):446–450.

Subramanian, S. V., Jones, K., Kaddour, A., and Krieger, N. (2009). Revisiting robinson: the perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, 38(2):342–360.

Surane, J. (2017). Goldman Sachs introduces real-time employee performance reviews. *Bloomberg Businessweek*. Retrieved from https://www.bloomberg.com/news/articles/2017-04-21/goldman-sachs-introduces-real-time-employee-performance-reviews.

Szymanski, K. and Harkins, S. G. (1987). Social loafing and self-evaluation with a social standard. *Journal of Personality and Social Psychology*, 53(5):891–897.

The Economist (2016). The measure of a man. *The Economist*, 418(8977):59.

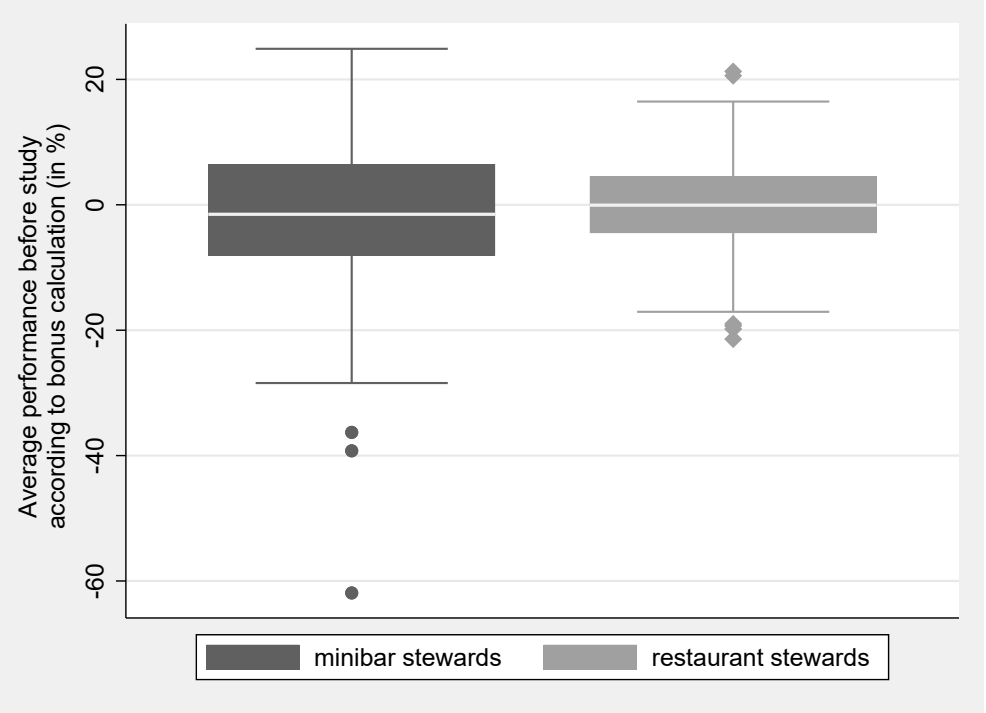
Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. (2018). Overcoming salience bias: How real-time feedback fosters resource conservation. *Management Science*, 64(3):1458–1476.

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

White, P. H., Kjelgaard, M. M., and Harkins, S. G. (1995). Testing the contribution of self-evaluation to goal-setting effects. *Journal of Personality and Social Psychology*, 69(1):69–79.

Yildirim, H. (2005). Contests with multiple rounds. *Games and Economic Behavior*, 51(1):213–227.

Appendix A Descriptive graphs



Notes: The performance per steward shown above is calculated using the weighted average of the monthly performance evaluations (see Section 1.3.1) over all months of the pre-study period.

FIGURE 1.4: Box plot of the pre-study steward performance

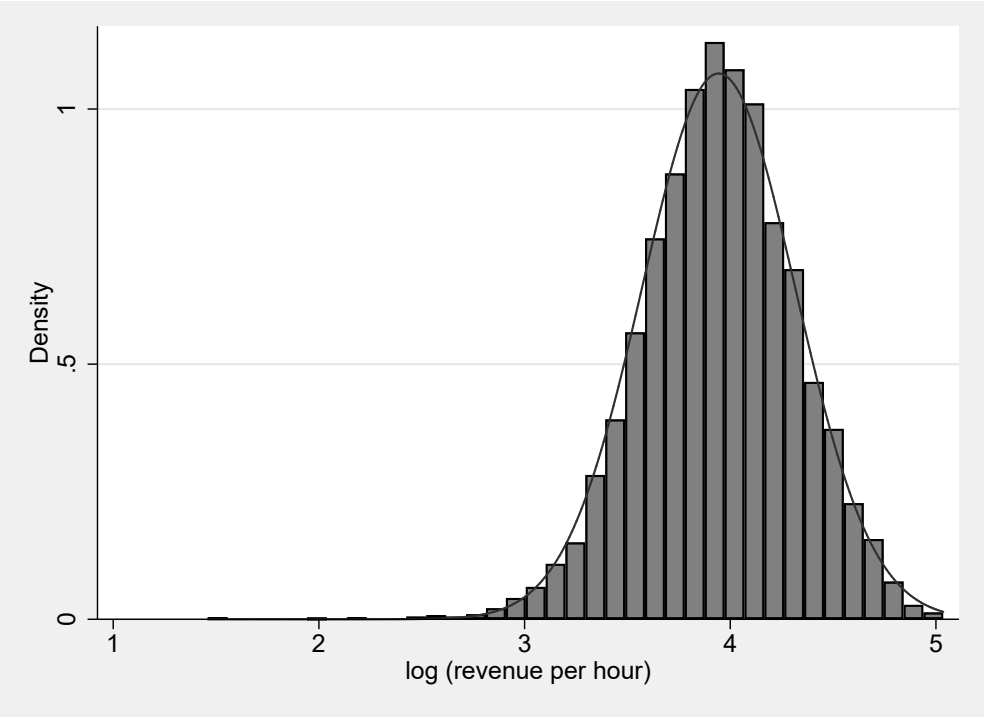
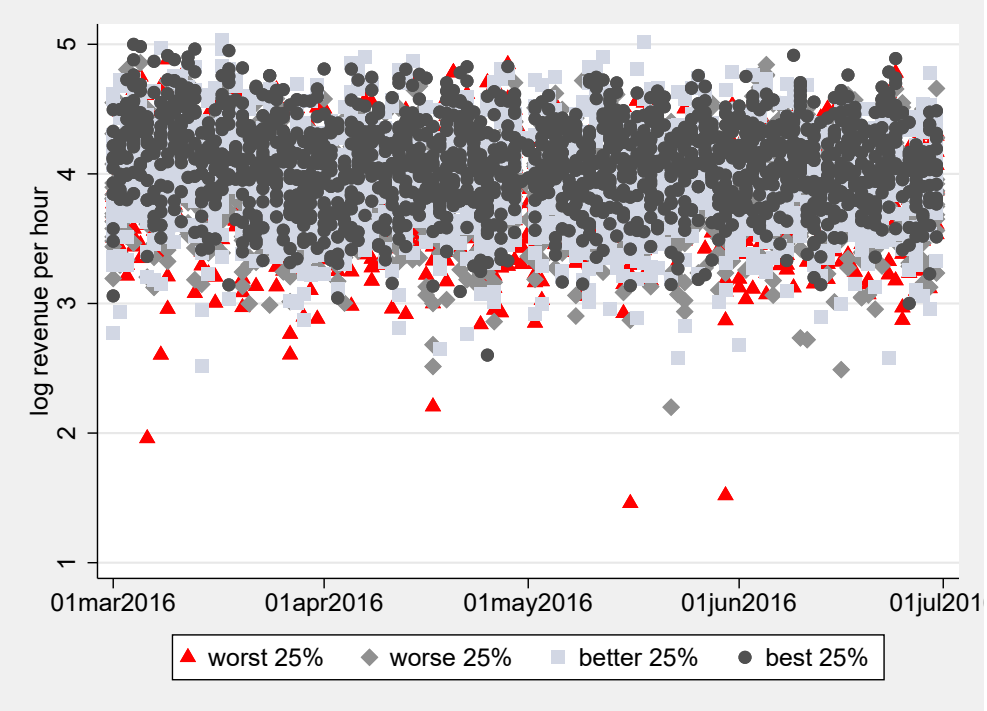


FIGURE 1.5: Histogram of the log revenue per hour in the study period



Notes: This figure illustrates the logarithmized revenue per hour for each service during the study period. The performance quartiles indicate whether the service was performed by one of the worst, worse, better, or best 25% of the minibar stewards. The quartiles refer to the stewards' average performance in the pre-study period.

FIGURE 1.6: Scatter plot of the log revenue per hour in the study period

Appendix B Robustness checks

TABLE 1.7: Pooled OLS regression: Log revenue per hour

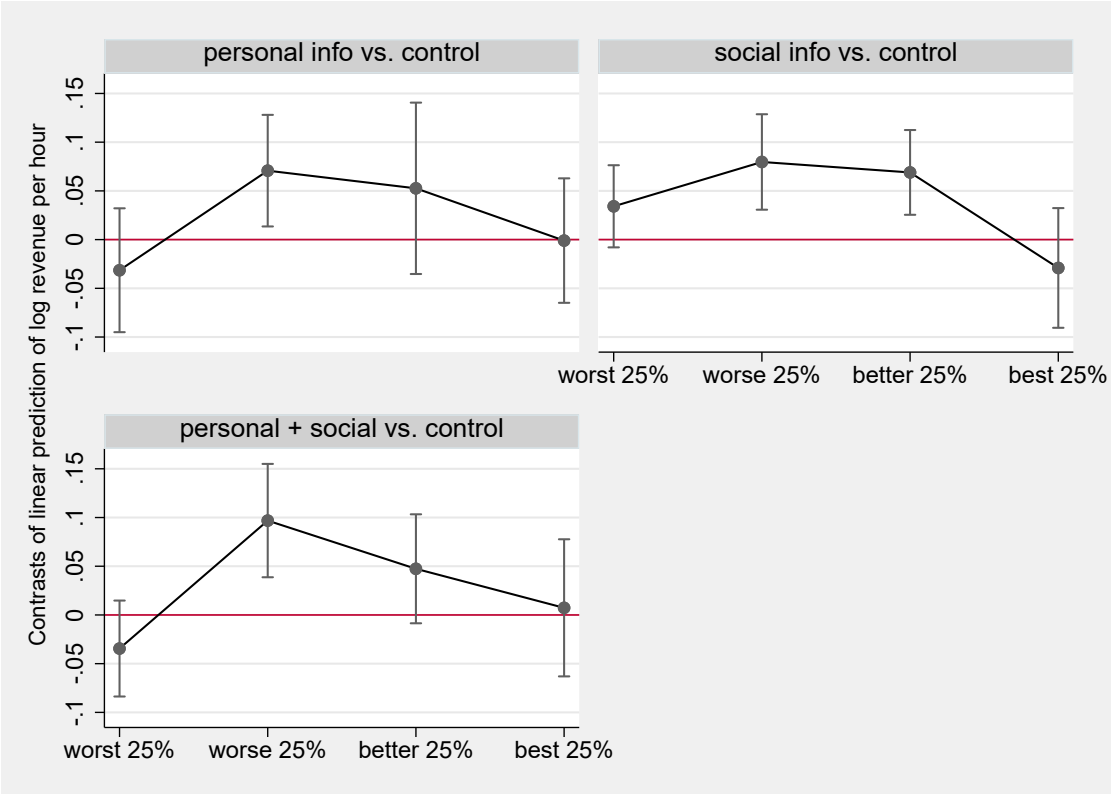
	(1) log revenue per hour
personal info	0.0174
	(0.0173)
social info	0.0329**
	(0.0129)
personal + social	0.0399**
	(0.0155)
performance before (in %)	0.0102***
	(0.0006)
worktime (in h)	-0.0141***
	(0.0032)
eating times (in %)	0.0076***
	(0.0007)
occupancy (in %)	0.0245***
	(0.0028)
occupancy^2	-0.0001***
	(0.0000)
1st class pass (in %)	0.0095***
	(0.0012)
no. stewards working	-0.7269***
	(0.1890)
event	0.1402***
	(0.0305)
weekend/holiday	0.1663***
	(0.0145)
shift type effects	Yes
city of shift start	Yes
month effects	Yes
R^2	0.376
Observations	6,149
N stewards	162

Notes: Pooled OLS regression of the logarithmized revenue per hour (per service). Robust standard errors clustered on the individual level are shown in parentheses. $personal\ info$, $social\ info$, and $personal + social$ are dummy indicators for the experimental treatments, whereas the control group is the reference category. All steward-, shift-, date-, and service-related control variables are included. See the discussion of Model 1.1 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 1.8: Difference-in-difference regressions: Log revenue per hour

	(1) DID RE regression	(2) DID FE regression	(3) DID FE without controls
personal info x study period	0.0221	0.0193	0.0090
	(0.0179)	(0.0178)	(0.0196)
social info x study period	0.0436***	0.0411***	0.0453***
	(0.0140)	(0.0140)	(0.0150)
personal + social x study period	0.0331*	0.0306*	0.0324*
	(0.0171)	(0.0173)	(0.0178)
performance before (in %)	0.0108***		
	(0.0004)		
worktime (in h)	-0.0035	-0.0044*	
	(0.0022)	(0.0024)	
eating times (in %)	0.0055***	0.0054***	
	(0.0005)	(0.0005)	
occupancy (in %)	0.0167***	0.0167***	
	(0.0005)	(0.0005)	
occupancy^2	-0.0000***	-0.0000***	
	(0.0000)	(0.0000)	
1st class pass (in %)	0.0071***	0.0068***	
	(0.0006)	(0.0006)	
no. stewards working	-0.8906***	-0.8759***	
	(0.1037)	(0.1063)	
event	0.1434***	0.1477***	
	(0.0232)	(0.0232)	
weekend/holiday	0.1229***	0.1213***	
	(0.0079)	(0.0079)	
shift type effects	Yes	Yes	No
city of shift start	Yes	Yes	No
month effects	Yes	Yes	No
sd (stewards)	0.034	0.153	0.188
sd (residual)	0.308	0.308	0.356
R^2 overall	0.388	0.281	0.000
Observations	32,928	32,928	33,064
N stewards	170	170	172

Notes: The table displays the treatment effects on the logarithmized revenue per hour (per service) in comparison to the pre-study period. Specification (1) includes random effects for stewards. Cluster-robust standard errors are shown in parentheses. In Specification (2) and (3) we use steward-fixed effects with single dummy variables for each person. The treatment group–study period interactions report the average change relative to the pre-study period in comparison to the control group (reference category). Specifications (1) and (2) include all steward-, shift-, date-, and service-related control variables. See the discussion of Model 1.1 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.



Notes: For each performance quartile (based on a steward's performance in the pre-study period), the graph shows the estimated marginal effects of a certain value of the treatment variable compared to the control group. We obtained these estimates from a difference-in-difference version of Model 1.2 with fixed effects for stewards. The error bars report the 95% confidence intervals.

FIGURE 1.7: Contrasts of predictive margins for the difference-in-difference estimates

Appendix C Feedback effects over time

TABLE 1.9: Random effects regression: Treatment effects over the study period

	(1) Interaction with study duration	(2) Interaction incl. square
study dur	0.0001	-0.0010
	(0.0003)	(0.0008)
personal info x study dur	-0.0005	-0.0023*
	(0.0004)	(0.0012)
social info x study dur	0.0000	-0.0005
	(0.0003)	(0.0011)
personal + social x study dur	-0.0001	-0.0017
	(0.0004)	(0.0015)
study dur^2		0.0000
		(0.0000)
personal info x study dur^2		0.0000
		(0.0000)
social info x study dur^2		0.0000
		(0.0000)
personal + social x study dur^2		0.0000
		(0.0000)
steward controls	Yes	Yes
shift controls	Yes	Yes
date controls	No	No
service controls	Yes	Yes
sd (stewards)	0.024	0.024
sd (residual)	0.291	0.291
R^2 overall	0.374	0.376
Observations	6,149	6,149
N stewards	162	162

Notes: The table displays the estimates of the logarithmized revenue per hour (per service), using a GLS regression with random effects for stewards. Robust standard errors clustered on the individual level are in parentheses. *personal info*, *social info*, and *personal+social* are dummy indicators for the experimental treatments, whereas the control group is the reference category. *study dur* is a continuous variable for the number of days since the start of the intervention. We do additionally control for weekends and public holidays but not for months. All other steward-, shift-, and service-related control variables are included. See the discussion of Model 1.1 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 1.10: Random effects regression: Treatment effects across months

	(1) interaction with day of month	(2) interaction with month periods
day	-0.0008	
	(0.0009)	
personal info x day	-0.0003	
	(0.0013)	
social info x day	-0.0004	
	(0.0013)	
personal + social x day	0.0006	
	(0.0014)	
middle		-0.0013
		(0.0209)
end		-0.0146
		(0.0199)
personal info x middle		-0.0019
		(0.0311)
personal info x end		-0.0159
		(0.0270)
social info x middle		-0.0021
		(0.0263)
social info x end		-0.0052
		(0.0270)
personal + social x middle		-0.0278
		(0.0299)
personal + social x end		-0.0027
		(0.0296)
steward controls	Yes	Yes
shift controls	Yes	Yes
date controls	Yes	Yes
service controls	Yes	Yes
sd (stewards)	0.020	0.021
sd (residual)	0.291	0.291
R^2 overall	0.376	0.377
Observations	6,149	6,149
N stewards	162	162

Notes: The table displays the estimates of the logarithmized revenue per hour (per service), using a GLS regression with random effects for stewards. Robust standard errors clustered on the individual level are in parentheses. $personal\ info$, $social\ info$, and $personal+social$ are dummy indicators for the experimental treatments, whereas the control group is the reference category. *day* is a continuous variable for the day of the month. *middle* and *end* are dummy variables indicating the middle and last 10 days of the month compared to the first 10 days that represent the reference group. All other steward-, shift-, date-, and service-related control variables are included. See the discussion of Model 1.1 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix D Immediate performance effects of real-time feedback

This additional section provides an in-depth analysis of employees' immediate reaction to the feedback messages during the service. By exploiting available data on single sales, we aim to investigate how comparative performance information affects immediate work performance directly after its release. These analyses should offer additional insights regarding the optimal timing of real-time performance feedback.

D.1 Introduction

Although regular performance feedback is a major trend in the business and private domains (consider, for instance, fitness and health trackers or social media likes), the *immediate* effects of of such information are hardly explored. Several authors have investigated the role of feedback frequency and immediacy in general (see Section 1.2.1) but did not address the direct impact of during-work feedback after its disclosure. Houde et al. (2013) study the effect of real-time feedback on electricity consumption in daytime. However, they do not consider the time of feedback release and its immediate impact on consumption.

Partially related to these analyses is the growing literature on interim performance feedback. In a principal–agent model with two periods, Lizzeri et al. (2002) show that the agent's total expected effort can be higher if his first-period outcome is revealed. Ludwig and Luenser (2008) find that intermediate feedback does not influence subjects' second-stage effort choices by itself but is conditional on their relative performance. Participants who lag tend to increase their second-stage effort, whereas those who lead tend to decrease it. In a similar setting, Aoyagi (2010) and Ederer (2010) suggest that the optimal disclosure policy depends on the agent's cost of effort function. Based on the assumption that agents know about their ability and that this knowledge enters the production function, Ederer (2010) further distinguishes between a beneficial "motivation effect" and an adverse "evaluation effect" of interim feedback. While interim information helps the agent in tailoring effort to his correct ability level, it also reveals how likely an agent is to win the tournament. This "evaluation effect" has a negative impact in the case of a large performance gap. Firms therefore face a fundamental trade-off when deciding whether to provide interim feedback. Goltsman and Mukherjee (2011) confirm this finding by showing that feedback disclosure policies that enhance final-stage effort may dampen incentives at the intermediate stage.

In an experimental study, Eriksson et al. (2009) find that information regarding the competitor's performance during a tournament leads to a performance increase for the

losing player if his score gap to is not too high. Like Eriksson et al. (2009), we empirically investigate the effect of intermediate feedback, but in our experiment the provided information is not novel. Stewards get to know their respective performance benchmarks (i.e., the personal and/or co-worker-related sales averages) at the beginning of their service and can access their current revenue at any time during work via the electronic till. We therefore investigate the immediate effect of interim feedback, which makes existing performance information more salient (also see Englmaier et al. 2017).

Our hypothesis is that salient during-work feedback leads to a performance increase directly after its release. Based on the literature around feedback and ability (see Section 1.2.3) and interim feedback disclosure, we further expect the temporal effect of the feedback message to depend on an employee's current performance–standard gap. We assume a positive, immediate impact on performance if the benchmark of the feedback message is perceived as difficult but attainable. However, we expect a negative, immediate performance impact if it becomes visible that an employee has already achieved the performance average or is highly likely to achieve it by the end of the service.

D.2 Empirical strategy

The original during-work feedback message, which was programmed to appear at a random time during the service, was not trackable. These messages were therefore reprogrammed on April 21, 2016, seven weeks after the start of the study. For the remaining 10 weeks of the experiment, the during-service messages were released according to a pre-defined timetable: two hours after the shift start (steward login) in the first week, three hours after login in the second week, four hours after login in the third week, and then again after two, three, and four hours in the subsequent weeks. Due to differing starting times of the shifts, the during-work messages on a certain date appeared at different times of the day. Figure 1.8 provides an illustration.

Based on this design, we are able to make a within-treatment comparison between those stewards who received the during-service feedback and those who did not (yet) obtain the message at a certain time after the shift start. More specifically, we compare the sales performance where the same steward on the same shift did receive the feedback message during the past 60 minutes and where he did not yet receive the during-work feedback. Because the control group in our intervention is expected to show a lower performance than the treatment groups across the whole service, we confine our analysis to the treatment groups. Including the control group might lead to an overestimation of the immediate performance effect.Figure 1.9 illustrates the within-treatment comparisons of the "message received" (grey) and "message not yet received" (cross-hatched) conditions.

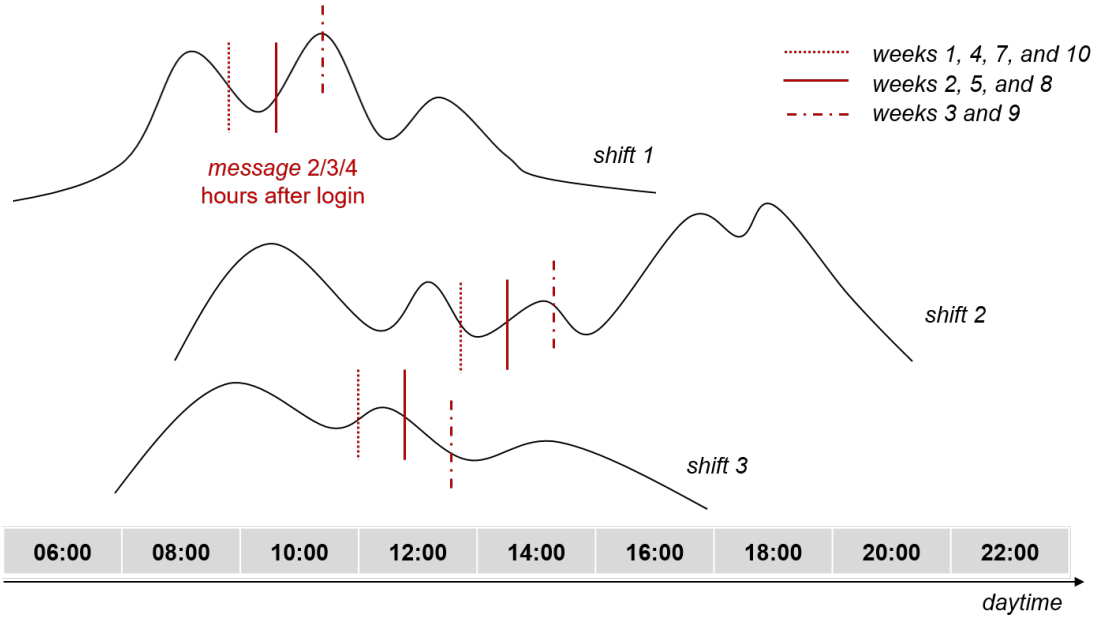


FIGURE 1.8: Illustration of the during-work feedback message

Because we are missing a clean comparison group for services where the message appeared after four hours, we confine our analysis to the third and fourth workin4g hour after login (as indicated by the black arrows).

To investigate time-related effects, we use data on the level of every single sale. Within each service (i.e., a shift performed by a certain steward on a certain date), we aggregate the single sales into sales per working hour, where the first working hour starts with the steward's login. Because we observe zero sales per hour for around 13% of our observations, we use the number of items sold per hour instead of the logarithmized sales revenue as our main outcome variable.[18] The number of items sold by steward i on service j in working hour t is estimated with the following Poisson regression model:

$$\begin{aligned} y_{i,j,t} = \beta_0 + \beta_1 message_{i,j,t} + \beta_2 working\ hour\ 4_t + \beta_3 break\ time_{j,t} \\ + \beta_4 occupancy_{j,t} + \beta daytime'_{j,t} + \beta day\ of\ week'_t j + \beta month'_j + \delta_i + \nu_j + \epsilon_{i,j,t}. \end{aligned} \tag{1.3}$$

Our main variable of interest is $message_{i,j,t}$, which is a dummy indicator for whether the feedback message was released to steward i on service j at the beginning of working hour t. We control for the duration a steward has been working on service j by the dummy variable *working hour 4* that indicates the fourth compared to the third working hour. The model further includes the break time of service j during hour t in minutes

[18] Logarithmic transformation would dismiss all zero values. Furthermore, there are many well-known approaches for estimating zero-inflated count data. Poisson regressions, in contrast to a Tobit model for example, also allow modeling multilevel structures.

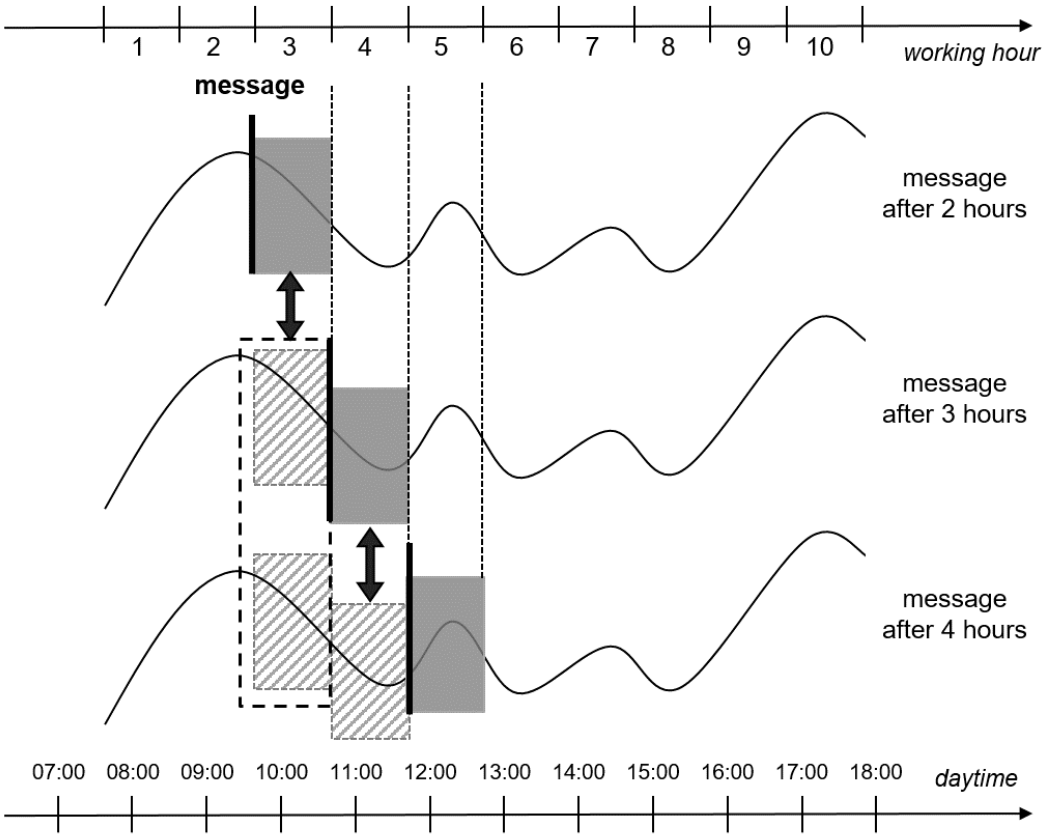


FIGURE 1.9: Model illustration

($break\ time_{j,t}$) and the passenger occupancy of service j during hour t in % ($occupancy_{j,t}$). We also take into account time- and date-related variables that could possibly influence a steward's sales performance during service j. These are dummy variables for the hour of the day (indicated by the vector $daytime'$), the day of the week ($day\ of\ week'$), and the month ($month'$). Finally, we use steward- and shift-fixed effects, indicated by δ_i and ν_j. $\epsilon_{i,j,t}$ captures any other unmodeled effects.

Our analysis includes all minibar services of the three treatment groups from April 21 until June 30, 2016.[19] Observations were excluded if the during-service message appeared during a break, during a change of trains, or if there was a train failure at any time during the service. As mentioned before, we further confine our analysis to the third and fourth working hours when some stewards already received the during-work feedback and others did not (see Figure 1.9). This leads to a total of 2,150 working hour observations (1,075 services with two working hours each). In 40% of these cases, the during-service message appeared two hours after login. The relative amount of observations for the three and four hours after login conditions are 28% and 32%.

[19] We did not use the control group in these within-treatment comparisons; however, the control condition could be added in further difference-in-difference analyses.

D.3 Results

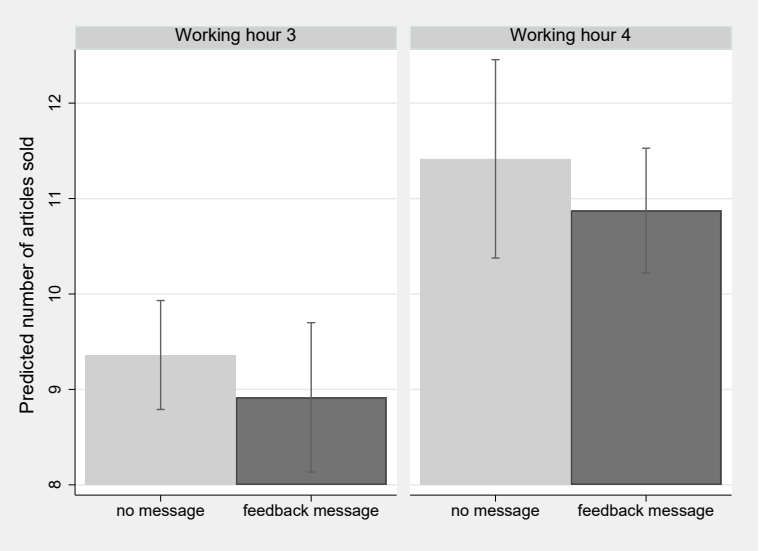
Table 1.11 shows the estimates for different specifications of Model 1.3. Taking all treatment groups together, we find a slightly negative but not significant effect of the during-service message on the number of items sold in the following 60 minutes. Stewards do not seem to sell more in working hour t right after the during-work message compared to the same steward on the same shift who did not yet receive a feedback message at the beginning of hour t. The coefficients for *break* and *occupancy* in Table 1.11 both point in the expected direction. The positive impact of the fourth in comparison to the third working hour may be explained by a clear peak in break time in working hour three. Stewards then possibly have additional energy or motivation in hour four. The right part of the table shows separate estimates of the full model (Specification 3) for the separate treatment groups. These results indicate a significant negative effect of the "social info" feedback during work on immediate sales performance. Within this treatment, the expected number of items sold is $(e^{0.169} - 1) * 100 = 15.5\%$ lower if a steward received the during-work message at the beginning of working hour t. Messages containing a personal performance benchmark, that is, the "personal" and "personal and social info" groups, do not seem to have an immediate impact.

Table 1.11: Poisson regression: Immediate feedback effect on items sold

	Within all treatment groups			Within single treatments		
	(1)	(2)	(3)	Personal Info	Social Info	Persona + Social
message	-0.071*	-0.022	-0.032	0.057	-0.169***	-0.019
	(0.040)	(0.035)	(0.035)	(0.071)	(0.060)	(0.057)
working hour 4	0.114***	0.182***	0.270***	0.255*	0.227*	0.249*
	(0.041)	(0.032)	(0.066)	(0.131)	(0.121)	(0.151)
break (in min)	-0.032***	-0.035***	-0.034***	-0.030***	-0.037***	-0.035***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.003)
occupancy (in %)	0.001	0.006***	0.008***	0.011***	0.006***	0.007***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
daytime FE	No	No	Yes	Yes	Yes	Yes
month FE	No	No	Yes	Yes	Yes	Yes
day of week FE	No	No	Yes	Yes	Yes	Yes
steward FE	No	Yes	Yes	Yes	Yes	Yes
shift FE	No	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.128	0.355	0.391	0.444	0.396	0.398
Observations	1,994	1,994	1,994	581	755	658
N Stewards	115	115	115	37	40	38

Notes: Poisson regression of the during-work feedback effect on the number of items sold in working hours three and four. Robust standard errors are shown in parentheses. *message* is a dummy variable that is equal to 1 if the feedback message appeared at the beginning of working hour t and 0 otherwise. *working hour 4* shows the general sales effect of the fourth compared to the third working hour. *break* and *occupancy* are continuous control variables for the break time and train occupancy rate in working hour t. Specification (3) and the estimates of the single treatment groups contain fixed effects (FE) for daytime (i.e., hour dummies), months, the days of the week, and for the shifts and stewards. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The results stay robust if we split the sales data into half-hour time frames and look at sales performance over the next 30 instead of the next 60 minutes after feedback release. Furthermore, separate estimates for working hours three and four reveal that the immediate performance effect is not significant, independent of whether the during-work message appeared two or three hours after service start (see Figure 1.10).



Notes: Poisson regression estimates (Model 1.3) for the number of items sold in working hours three and four. The error bars report the 95% confidence intervals.

FIGURE 1.10: Predicted number of items sold

To further understand this outcome, we also consider a steward's current sales performance at the time of feedback release.[20] We measure current performance by the remaining revenue a steward has to generate per hour to achieve the performance benchmark of his feedback message. In terms of the "personal info" ("social info") treatment group, this benchmark is the personal (total) average revenue of the (all) steward(s) on the same shift during the past 30 days. For the "personal and social info" condition, we chose the social average to calculate the current performance measure.[21] By taking into account the remaining working time, we ensure that the performance measure is independent of the time when the during-service message appeared (i.e., two, three, or four hours after login). If a steward has already achieved the benchmark at the time of message release, the current performance measure becomes negative. However, this only occurs in 3% of the cases. The median of the remaining revenue per hour lies at 39.5 CHF.

[20] While we were looking at the interaction with a steward's general ability in Section 1.4.2, the focus here is on the sales that a steward hitherto generated on his *current* service.

[21] As previous analyses revealed, personal performance information has a lower impact on performance (see Section 1.4.1).

Table 1.12 shows the estimates of Model 1.3, including the interaction with a steward's current performance. *rev to go* indicates the remaining revenue per hour before steward i achieves his performance benchmark on service j. Taking all treatment groups together, we observe a significant negative impact of the during-work message at the time of benchmark achievement when *rev to go* is 0 (see negative coefficient of *message*). However, the during-work feedback effect becomes more positive as the deviation from the feedback benchmark increases, that is, when *rev to go* becomes larger (see positive coefficient of *message* x *rev to go*). This interaction effect is even stronger when using alternative model specifications, such as a negative binomial regression and a mixed Poisson model with random effects for each service (see Hedeker and Gibbons 2006, pp. 239-256, Atkins et al. 2013).[22]

[22]We also obtain a similar interaction effect if we use quartile dummies for the stewards' current performance instead of the continuous *rev to go* measure as mediators. Including the quadratic term *rev to go*^{2} in addition to *rev to go* has no significant impact on the results. We therefore assume that employees do not slack off, even if their present performance is very poor. All additional results are available on request.

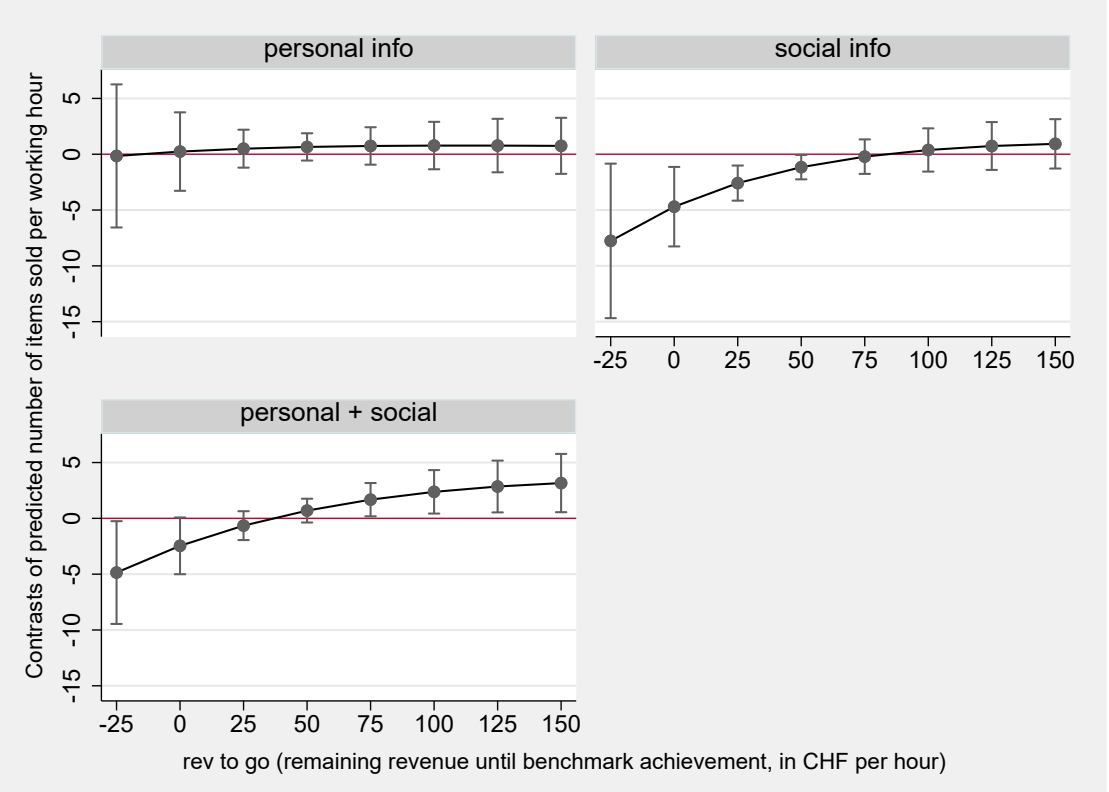
TABLE 1.12: Poisson regression: Interaction effect with current performance

	Within all treatment groups			Within single treatments		
	(1)	(2)	(3)	Personal Info	Social Info	Personal + Social
message	-0.201**	-0.216***	-0.166**	0.006	-0.434***	-0.193*
	(0.081)	(0.073)	(0.068)	(0.134)	(0.126)	(0.105)
rev to go (CHF per h)	-0.003*	-0.011***	-0.009***	-0.009***	-0.014***	-0.007***
	(0.002)	(0.001)	(0.001)	(0.003)	(0.002)	(0.002)
message x rev to go (CHF per h)	0.003*	0.005***	0.003**	0.001	0.005**	0.005**
	(0.002)	(0.002)	(0.001)	(0.003)	(0.003)	(0.002)
working hour 4	0.112***	0.136***	0.199***	0.239*	0.096	0.182
	(0.042)	(0.032)	(0.065)	(0.131)	(0.119)	(0.147)
break (in min)	-0.032***	-0.035***	-0.034***	-0.030***	-0.037***	-0.035***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)
occupancy (in %)	0.000	0.005***	0.006***	0.009***	0.004*	0.006***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
day time FE	No	No	Yes	Yes	Yes	Yes
month FE	No	No	Yes	Yes	Yes	Yes
day of week FE	No	No	Yes	Yes	Yes	Yes
steward FE	No	Yes	Yes	Yes	Yes	Yes
shift FE	No	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.130	0.369	0.400	0.454	0.413	0.402
Observations	1994	1994	1994	581	755	658
N Stewards	115	115	115	37	40	38

Notes: Poisson regression of the during-work feedback–current performance interaction. Dependent variable is the number of items sold in working hours three and four. Robust standard errors are shown in parentheses. *message* is a dummy variable that is equal to 1 if the feedback message appeared at the beginning of working hour t and 0 otherwise. *rev to go* indicates the remaining revenue per hour before steward i achieves the benchmark shown in his feedback message. *working hour 4* shows the general sales effect of the fourth compared to the third working hour. *break* and *occupancy* are continuous control variables for the break time and train occupancy rate in working hour t. Specification (3) and the estimates of the single treatment groups contain fixed effects (FE) for daytime (i.e., hour dummies), months, the days of the week, and for the shifts and stewards. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As the results in the right part of Table 1.12 suggest, the feedback effect is particularly sensitive to a steward's current performance in the "social info" treatment. Here, the estimated decrease in subsequent sales is $(e^{0.434}-1)*100 = 32.2\%$ if the feedback message appears *after* an employee has reached the co-worker-related performance average (i.e., when *rev to go* is 0). Congruent with previous findings of this paper, stewards in the "personal info" condition do not show any response to the during-work feedback, independent of their current sales revenue.

Figure 1.11 provides the predictive margins of the regression analyses for each treatment group. The graphs show the predicted number of items sold in hour t when a steward just received the during-work message compared to the case when he did not (yet) receive the during-work feedback. The immediate feedback effect in the "personal and social info" group becomes positive for stewards who still have to earn more than 36.2 CHF per hour during the remaining time of their service. Within the "social info" group, the respective turning point lies at 82.8 CHF per hour, implying that the immediate performance impact is positive in only 1–2% of the cases.



Notes: For each treatment group, the graph shows the estimated marginal effect (Model 1.3) of the during-work feedback for different values of a steward's current performance. Current performance is decreasing as the remaining revenue, *rev to go*, increases from left to right. The error bars indicate the 95% confidence intervals.

FIGURE 1.11: Contrasts of predictive margins for the feedback–performance interaction

Table 1.13 further shows the feedback–performance interactions for working hours three and four separately. Interestingly, the interaction effect seems to be driven by working

hour four, that is, for feedback messages that were released later on in the working day. Although we are not able to track this trend further, this result indicates that the immediate effect may be more sensitive to a worker's current performance if feedback is revealed towards the end of the service. Earlier on in the working day, in contrast, the gap to the performance benchmark seems to be less influential. It is certainly conceivable that relative feedback which is disclosed at a later stage of the task causes greater pressure to perform than early feedback. However, whether the feedback–performance interaction indeed depends on the timing of feedback disclosure needs further investigation in future studies.

TABLE 1.13: Separate Poisson regressions for the number of items sold

	Working hour 3	Working hour 4
message	-0.116	-0.357***
	(0.083)	(0.099)
rev to go (CHF per h)	-0.012***	-0.009***
	(0.002)	(0.002)
message x rev to go (CHF per h)	0.002	0.006***
	(0.002)	(0.002)
break (in min)	-0.032***	-0.024***
	(0.005)	(0.007)
occupancy (in %)	0.006***	0.000
	(0.001)	(0.001)
daytime FE	Yes	Yes
month FE	Yes	Yes
day of week FE	Yes	Yes
steward FE	Yes	Yes
shift FE	Yes	Yes
Pseudo R^2	0.521	0.537
Observations	997	597
N Stewards	115	108

Notes: Separate Poisson regression of the during-work feedback–current performance interaction in working hours three and four. Dependent variable is the number of items sold per working hour. Robust standard errors are shown in parentheses. *message* is a dummy variable that is equal to 1 if the feedback message appeared at the beginning of working hour t and 0 otherwise. *rev to go* indicates the remaining revenue per hour before steward i achieves the benchmark shown in his feedback message. The estimates include fixed effects (FE) for daytime (hour dummies), months, the day of the week, and for each shift and steward. See the discussion of Equation 1.3 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D.4 Discussion

While performance feedback is one of the most extensively studied fields in behavioral economics, there is still little knowledge about how feedback affects performance immediately after its release and over the duration of a task. Our analyses shed some light on this question. Against our hypothesis, the results do not confirm an immediate

positive effect of during-work feedback on subsequent sales performance. However, employees still react to salient performance information directly after its release. Workers who perform far below the social average tend to be immediately motivated by messages that include co-worker-related benchmarks. Feedback underlining that the social performance benchmark is likely to be reached, however, has a significant negative impact on immediate performance. This especially applies for well-performing workers who only receive social feedback and no personal performance standard they can additionally compete against. Overall, we find that the immediate effectiveness of feedback containing social (or social and personal) performance information crucially depends on a worker's performance at the time of feedback release. Personal performance information alone seems to have no effect on immediate sales, independent of an employee's current position.

These results are consistent with the literature around interim performance feedback, suggesting that an agent's reaction to peer-related feedback depends on his relative performance (e.g., Ludwig and Luenser 2008, Ederer 2010, Goltsman and Mukherjee 2011). It also confirms our previous finding that performance ability is an important mediator for the effectiveness of real-time feedback. Our analyses now provide a first indication that this interaction effect also holds for the immediate impact of feedback throughout the day.

Our findings also show preliminary evidence for a heterogeneous course of the feedback–performance interaction. It appears that an employee's current performance has a stronger influence if feedback is disclosed towards the end of the task rather than at the beginning. Conversely, the immediate reaction to feedback seems to be less affected by an employee's present level of attainment when feedback is provided at an early stage. We explain this result with a lower urgency of effort adjustment if performance gaps become salient early on.

Although the validity of these ideas needs to be tested in future studies, our findings allow some preliminary suggestions for practice. First, despite the overall positive effects of timely co-worker-related performance information (see Section 1.4.1), making this type of feedback salient during work is not a general means for immediate, short-term improvements. Our analyses rather suggest that social performance information during work should be provided selectively for interim poor performers to prevent potential negative effects. Furthermore, if the selective provision of feedback is not feasible, it may be reasonable for companies to disclose vague performance information. Partial disclosure or vague feedback instead of full revelation of interim performance have been proposed as optimal strategies in previous studies (e.g., Hannan et al. 2008, Goltsman and Mukherjee 2011). Companies may also implement some kind of partial disclosure

policy by providing feedback at an early stage of the task when the final outcome is still indefinite.

Looking ahead, the immediate effect of real-time feedback, its connection with current performance levels, and the role of different times of feedback disclosure require further research. In particular, the present analyses are limited to the specific characteristics of our field setting, where, for example, during-work feedback is not novel but is only made more salient to employees. Likewise, the incentive scheme differs from existing literature on intermediate performance information in tournaments (e.g., Casas-Arce and Martínez-Jerez 2009). Additional insights into the questions previously mentioned would be beneficial for the optimal timing of feedback messages in practice. This is especially relevant, as information technology provides ever wider options for customized feedback systems in commercial and private spheres.

Essay 2: Choose to reuse! The effect of action-close reminders on pro-environmental behavior

Andrea Essl, Angela Steffen, Martin Staehle [*]

Abstract

Individuals regularly underinvest in activities that provide future, collective benefits but require immediate effort. We study whether simple reminders are an effective means to promote investments in environmental protection, namely the prevention of plastic waste. In a field experiment with a Swiss food box provider, customers received weekly reminders highlighting the option to return plastic bags for reuse. We find that reminders are highly effective in reducing plastic waste as they increase customers' plastic bag return rate by up to 83%. This effect persists over the intervention period and beyond. Importantly, reminders are most effective if they are action-close, that is, when they raise attention to the issue in close proximity to the decision-making situation. Our study provides insights into the attentional mechanisms underlying reminder effects and highlights action-closeness as an opportunity to effectively implement reminders in practice.

Keywords: reminder, environmental behavior, plastic waste, limited attention, randomized controlled trial, decision-making, habit formation

[*]Andrea Essl: andrea.essl@iop.unibe.ch, Institute for Organization and Human Resource Management, University of Bern; Angela Steffen: angela.steffen@iop.unibe.ch, Institute for Organization and Human Resource Management, University of Bern; Martin Staehle: martin_staehle@hotmail.com. We would like to thank the partner association, in particular Fredy Gmuer, who made this study possible. We are also grateful to Christian Zehnder, Holger Herz, Robert Dur, and the participants of the CUSO workshop 2018 and the internal research seminar at the University of Bern 2017. All errors are ours. This study is registered in the American Economic Association's registry for randomized controlled trials with the unique identifying number: AEARCTR-0002523.

2.1 Introduction

Plastic waste has become a major environmental issue for humanity (UNEP 2018). In 2015, the total quantity of plastics ever produced amounted to 8,300 million metric tons, 60% of which had accumulated in landfills and the environment (Geyer et al. 2017). Packaging dominates this flow of plastic waste (Dahlbo et al. 2018). In fact, plastic packaging is often discarded after the first use and accounts for nearly 50% of all plastic waste globally (UNEP 2018). Environmentally friendly alternatives to wasting plastics are source reduction, product reuse, and recycling (Stein 1992, Hopewell et al. 2009). Plastic bags, for example, can be reused several times, and the simple action of reuse yields significant environmental benefits (Bisinella et al. 2018). However, reducing the burden of plastic waste by product reuse requires behavioral change (see MacArthur 2017).

Even if most people have an intention to behave in an environmentally friendly way, they often fail to do so in their daily lives. Reusing plastics, for instance, is often associated with immediate effort and laborious and sometimes dirty activities (Barr 2002). The subsequent environmental benefit is collectivized and delayed. Research has frequently shown that individuals do not keep up on their intentions with respect to such investment activities (Charness and Gneezy 2009, Karlan et al. 2016, Calzolari and Nardotto 2017). One explanation for underinvestment in pro-environmental behavior and other desirable actions may be limited attention. In accordance with the salience bias (Kahneman et al. 1982, Bordalo et al. 2012, Tiefenbeck et al. 2018), individuals with limited attention have a tendency to focus on salient aspects of behavior (such as the immediate costs of an effort) and ignore less-salient implications (such as future benefits).

Reminders are one way to refocus attention on actions that may otherwise be forgotten. More specifically, reminders can bridge the attentional gap between intention and action, because they enhance the processing of future-oriented aspects and bring essential future implications of behavior to the top of the mind (Borgstede and Andersson 2010, Taubinsky 2013, Karlan et al. 2016). Research further suggests that attention is attracted by salient aspects of the choice environment (Bordalo et al. 2012, Szilagyi and Adams 2012, Karlan et al. 2016). Action-close reminders are memory aids that catch people's attention in the situation and at the time of the desired behavioral change. Such reminders have been shown to effectively promote investment activities in various contexts (e.g., Luyben 1980, Jacobs and Bailey 1982, Krendl et al. 1992, Werner et al. 1998, Shearer et al. 2017).

In this paper, we examine how far simple reminders, and particularly action-close reminders, can enhance pro-environmental behavior and increase investments in plastic waste reduction. To analyse this question, we conduct a randomized controlled trial with

287 customers of a Swiss agricultural association that delivers weekly food boxes with different vegetables wrapped in plastic bags. The customers have the option to return these bags to the association for reuse. Prior to our study, however, return rates were very low. As an intervention, we add a weekly reminder to the customers' food boxes. These reminders are either provided in the form of a flyer, which is separately added to the food box, or in the form of a sticker, which is directly attached to one of the plastic bags in the food box. Both reminders are designed to revamp attention toward the return option for plastic bags. We therefore conjecture that the flyer and sticker reminders increase customers' return rates of the plastic bags in comparison to a control group with no reminders. In line with previous literature that stresses the power of habits (Volpp et al. 2008, Charness and Gneezy 2009, Acland and Levy 2015, Shearer et al. 2017), we further hypothesize that the reminder effect has the potential to persist throughout the intervention period and beyond. We present the sticker reminders in close proximity to the decision as to whether the plastic bags should be reused or discarded. Therefore, the action-close sticker reminders should be able to refocus customers' limited attention to the desired return option to a greater extent than conventional reminders (Austin et al. 1993). We thus hypothesize that plastic bags marked with an action-close sticker reminder are more likely to be returned than unmarked plastic bags in the sticker and flyer treatments.

Overall, we find a statistically significant and quantitatively large effect for both the flyer and sticker reminders on plastic waste reduction. Relative to the pre-intervention period, the increase in return rates of plastic bags during the intervention was up to 83% higher in the reminder treatments than in the control group. Remarkably, the impacts of the flyer and the sticker reminders were stable during the five-week intervention period, indicating that repeated reminders do not lose their effectiveness. We also observe a significant positive effect for both reminder treatments in the post-intervention period, suggesting that the behavioral change induced by the reminders is, to a certain degree, persistent over time. In line with our hypotheses, we further find that a reminder's proximity to action significantly improves its effectiveness. The probability of returning a bag with an action-close sticker reminder attached was up to 58% higher than the probability of returning an unmarked bag in either of the treatment groups. This finding provides new evidence that action-close reminders may be more effective than conventional reminders. With respect to unmarked bags, we find no difference between the flyer and sticker treatments, indicating that both forms of reminders are similarly effective when they are distanced from the decision situation.

To the best of our knowledge, we are one of the first to investigate the role of a reminder's proximity to action. Previous studies solely focus on the effect of point-of-decision prompts to encourage different types of investment activities without comparing their

impact to conventional reminders (Luyben 1980, Werner et al. 1998, Russell et al. 1999, Sussman and Gifford 2012, Allais et al. 2017, Shearer et al. 2017). Austin et al. (1993), as an exemption, show that sign prompts improve recycling behavior when they are close to the point of decision. Together with the visual reminder, however, they also vary the positioning of receptacles and, thus, the effort associated with the recycling activity. In this paper, in contrast, we directly compare the effectiveness of action-close and conventional reminders. By showing that a reminder's proximity to action is important by itself, we provide new evidence that action-close reminders can more effectively bridge limited attention.

Besides extending behavioral research on reminder effects (e.g., Apesteguia et al. 2013, Altmann and Traxler 2014, Karlan et al. 2016, Calzolari and Nardotto 2017), our paper belongs to a growing number of empirical studies that examine pro-environmental interventions. Reduced individual showering times (Attari et al. 2010, Attari 2014, Tiefenbeck et al. 2018), general household energy savings (Rea et al. 1987, Schultz et al. 2007, Werner et al. 2012, Allcott and Rogers 2014), and recycling (Miafodzyeva and Brandt 2013, Shearer et al. 2017) are only few examples. Our paper differs form this work mostly by examining the effectiveness of action-close reminders in a new environmental setting, namely in the area of plastic waste reduction through reuse.

From a practical perspective, our findings suggest reminders as an effective, low-cost, and easy-to-implement option to encourage pro-environmental behavior. Sustainable behavioral change may therefore not require complex informational messaging or feedback. Furthermore, our results provide important implications for the effective implementation of reminders in practice. Reminders should optimally be issued at the time and in the situation when action takes place. Last but not least, the behavioral change in our setting may also benefit the customers of the food box provider and the agricultural association itself. The reminders effectively support the transfer of customers' environmentally friendly intention into action. This is likely to enhance their satisfaction with the food box offer.

The remainder of this paper is structured as follows. Section 2.2 provides detailed information on the field setting, outlines the experimental design, and presents the sample characteristics and randomization checks. Section 2.3 lays out our experimental results, and Section 2.4 provides concluding remarks.

2.2 Field experiment

2.2.1 Field setting

We conducted our field experiment in cooperation with a Swiss agricultural association that offers weekly compilations of organic farm products. These food boxes can be bought through an annual subscription that is available for three different types (meat, vegetarian, or vegan) and in two sizes (large and small). The food baskets contain regional, organic, and seasonal vegetables, and depending on the food box type, eggs, meat, or other farm products. The annual subscription includes 48 deliveries and four (self-determined) holiday weeks per year for which the delivery is suspended. In terms of distribution, the baskets are labelled with the company logo and customer name and are delivered every week to 11 depots in the city of Bern. Customers can then pick up their food boxes from one of these depots.

Since many vegetables are pre-portioned (e.g., carrots), loose (e.g., baby spinach) or wet (e.g., fresh salad), about 60% of the products are wrapped in plastic bags. For the partner association, reusing these plastic bags in the delivery process is one of the most sustainable and financially viable packaging options. The literature similarly suggests that plastic bags can be reused several times and have less environmental impact than, for instance, organic cotton bags (Bisinella et al. 2018).[23] New subscribers are therefore explicitly encouraged in the welcome letter to return the plastic bags for reuse. In terms of sustainability, plastic bags should ideally be reused as many times as possible (Bisinella et al. 2018). Although we expect that food box subscribers share a common intention to behave in an environmentally friendly way, the return rate of the plastic bags prior to the intervention used to be very low (16.7%).

2.2.2 Experimental design and procedure

In this field experiment, we sought to examine whether and how simple reminders can encourage customers to return plastic bags for reuse. We used a between-subject design with two experimental treatments and a control group.[24] In the control group, no reminder was in place. In the flyer treatment, a conventional flyer was added to the food

[23] Organic cotton bags have to be used at least 149 times to offset their climate impact; this is compared to 43 times for regular paper bags and once for low-density polyethylene (LDPE) plastic bags (Bisinella et al. 2018).

[24] The experimental details were pre-registered on the American Economic Association's registry for randomized controlled trials with the unique identifying number AEARCTR-0002523.

box, reminding customers to return the plastic bags for reuse.[25] In the sticker treatment, the reminder was directly attached to one of the plastic bags (see Figure 2.1). The provision of only one reminder per delivery week in each treatment should ensure the comparability of the experimental groups and, at the same time, allow the evaluation of the action-closeness effect.

FIGURE 2.1: Reminder treatments

Customers were randomly assigned to the flyer treatment, sticker treatment, or control group. To ensure an equal distribution in terms of customers' geographical locations and food box types, we stratified the sample according to the depots, basked types, and basket sizes. The reminders in both experimental treatments were equal in terms of content, layout, and size. Several studies indicate that simplicity, noticeability, and clearness may improve the effectiveness of visual reminders (e.g., Williams et al. 1989, Kline and Beitel 2016). Furthermore, research suggests that adding a picture that emphasizes the message may support its impact (e.g. Werner et al. 1998, Jae et al. 2008, Roberts et al. 2009). We took these factors into account during the design process. Both types of reminders contained the following information: "Please return the plastic bags. They can be reused." The reuse symbol was used to support the written information (see Figure 2.8 in Appendix A).

[25]The flyer was not directly attached to the food box but put inside so customers in the depots could not see them from the outside. We do not assume confounding effects from personal communication, since customers pick up their baskets at individual times and in semi-public places, such as staircases or storage rooms, where people usually do not linger.

The intervention took place for five delivery weeks from October 18 until November 15, 2017. During this period, treated customers received weekly reminders either in the form of a sticker attached to one of their plastic bags or in the form of a flyer. Beyond the intervention period, we tracked customers' returning behavior for two delivery weeks before (pre-intervention period) and four delivery weeks after (post-intervention period) the intervention. Plastic bags were therefore tracked for eleven weeks in total, between October 4 and December 20, 2017. Customers did not always return the empty food boxes and the corresponding plastic bags from the previous week when they picked up their new food boxes. Since the time of delivery and the time of return of the plastic bags may fall more than one week apart, we counted the returns for an additional four weeks after the post-invention period, until January 10, 2018. The timeline of the experiment is illustrated in Figure 2.2.

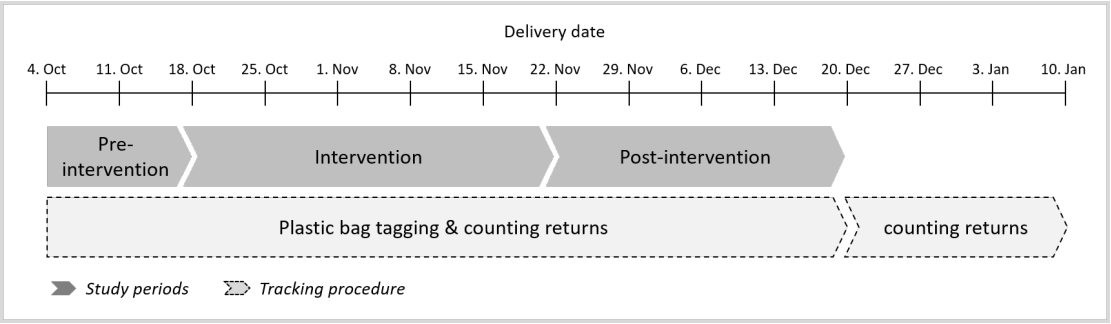


FIGURE 2.2: Timeline of the experiment

The main challenge during the data collection process was to track the number of returned plastic bags on an individual level. We therefore tagged every plastic bag with an almost invisible ID label, indicating a unique identification number for the customer and the current delivery week (see Figure 2.9 in Appendix A). This procedure allowed us to gather individual customer data on plastic bags delivered and returned for each week during the study period. To avoid confounds, the plastic bags that were returned during the experiment were not reused directly but collected by the experimenters and reintroduced by the organization after the end of the study.

To investigate and compare the effects of the flyer and sticker reminders in general (see Sections 2.3.1 and 2.3.2), we employ the return rate per customer as our main outcome measure. This variable is calculated using the number of returned plastic bags from each delivery week divided by the total number of delivered bags per week.[26] To evaluate the effect of a reminder's proximity to action, we compare the probability of return for

[26]Note that we analyse the returned plastic bags per delivery week, irrespectively of the time of their return. This allows us to clearly attribute the plastic bag observations to the pre-, during-, or post-intervention period.

plastic bags marked with action-close sticker reminders to the probability of return for bags without attached sticker reminders in the flyer, sicker, and control conditions. The comparison of marked plastic bags in the sticker treatment with unmarked plastic bags in the flyer treatment allows us to identify the superior effect of action-close compared to conventional reminders. The comparison of marked and unmarked plastic bags within the sticker treatment further permits us to explore a potential spillover effect of action-close reminders to those decision situations where no action-close reminder is present. Our main dependent variable in these analyses is a dummy variable indicating whether the plastic bag was returned for reuse or not (see Section 2.3.3). In addition to the outcome measures, we use data on whether the basket itself was returned or not, the delivery week, and the customers' food box types, sizes, and depots as control variables (see Section 2.2.3).

2.2.3 Sample characteristics and randomization checks

337 customers of the agricultural association participated in our experiment. The return rate of plastic bags in the pre-intervention period could not be observed for 50 customers who were on holiday in either of the pre-intervention weeks.[27] As the return rate in the pre-intervention period is essential for our analyses (see Section 2.3.1), we consider the remaining 287 customers as our final sample.[28] Out of these 287 customers, 93 received flyer reminders, 96 received sticker reminders, and 98 received no reminders. Each basket contained an average of 5.2 plastic bags with a standard deviation of 1.16. This leads to a total of 7,760 plastic bag observations in our final data set. Figure 2.3 shows the number of observations on both the individual- and bag-level of analysis.

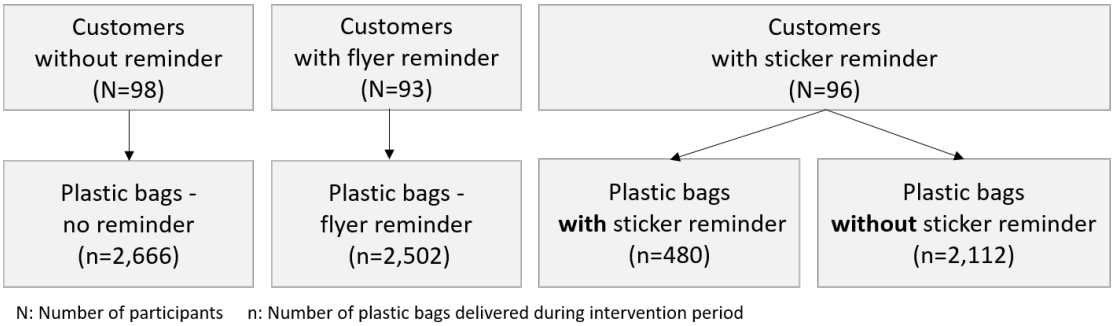


FIGURE 2.3: Levels of analysis

[27] Holiday weeks during the the intervention period were treated as missing observations for the respective customers.

[28] The results for the total sample are robust and available upon request.

Table 2.1 further provides the observed customer characteristics and the average return rates in the pre-intervention period for the whole study sample and for each treatment group separately.

TABLE 2.1: Sample characteristics and randomization checks

	Sample n= 287	Control n=98	Flyer n=93	Sticker n=96	p-value
Small box	0.92	0.95	0.90	0.92	0.436
Big box	0.08	0.05	0.1	0.09	0.436
Meat box	0.30	0.32	0.30	0.28	0.867
Veggie box	0.49	0.47	0.51	0.50	0.865
Vegan box	0.21	0.21	0.19	0.22	0.903
Depot 1	0.13	0.15	0.13	0.10	0.601
Depot 2	0.05	0.04	0.05	0.05	0.903
Depot 3	0.04	0.02	0.05	0.04	0.499
Depot 4	0.04	0.04	0.05	0.03	0.745
Depot 5	0.02	0.02	0.02	0.02	0.999
Depot 6	0.11	0.09	0.13	0.10	0.705
Depot 7	0.15	0.16	0.14	0.16	0.900
Depot 8	0.01	0.01	0.01	0.01	0.999
Depot 9	0.13	0.13	0.13	0.14	0.992
Depot 10	0.12	0.11	0.10	0.15	0.569
Depot 11	0.20	0.21	0.18	0.20	0.862
Holiday weeks (intervention)	0.43	0.54	0.44	0.31	0.132
	(0.82)	(0.94)	(0.81)	(0.65)	
Basket returned (intervention)	6.51	6.83	6.65	6.06	0.174
	(2.94)	(2.55)	(2.79)	(3.4)	
Return rate (pre-intervention)	0.17	0.14	0.18	0.18	0.540
	(0.25)	(0.25)	(0.23)	(0.26)	

Notes: The table reports means and standard deviations for continuous variables and percentage frequencies for categorical variables for the full sample and for each treatment group individually. Standard deviations are given in parentheses. For categorical variables, the p-value in the last column was obtained from a $\tilde{\chi}^2$-test across all experimental groups. For continuous variables, the p-value was obtained from an F-test.

Consistent with the randomization procedure, the average customer does not differ in terms of observed characteristics across treatments (F-test). In the pre-intervention period, treated customers returned slightly more plastic bags than those in the control group (18% in each reminder group versus 14% in the control group). However, according to the F-test, this difference is not significant. Since a balanced pre-intervention return rate is crucial for our analyses, we additionally conduct a pairwise comparison. The two-sided t-tests, however, do not reject the null hypothesis of an equal mean for the treatments and the control group (p=0.378 flyer vs. control, p=0.317 sticker vs. control, p=0.864 sticker vs. flyer).

2.3 Results

2.3.1 The impact of reminders on return behavior

To analyse the effect of reminders on return behavior, we compare the return rates across treatments in the pre-, post- and intervention periods. Table 2.2 reports the average return rates for each treatment group in each period.

TABLE 2.2: Average return rates over the study periods

	Pre-intervention n=287	Intervention n=287	Post-intervention n=286	Difference intervention–pre-intervention
Control (n=98)	0.14	0.14	0.12	0.00
	(0.25)	(0.23)	(0.20)	(0.34)
Flyer (n=93)	0.18	0.33***	0.26***	0.15
	(0.23)	(0.29)	(0.27)	(0.37)
Sticker (n=96)	0.18	0.32***	0.22***	0.14
	(0.26)	(0.29)	(0.25)	(0.39)

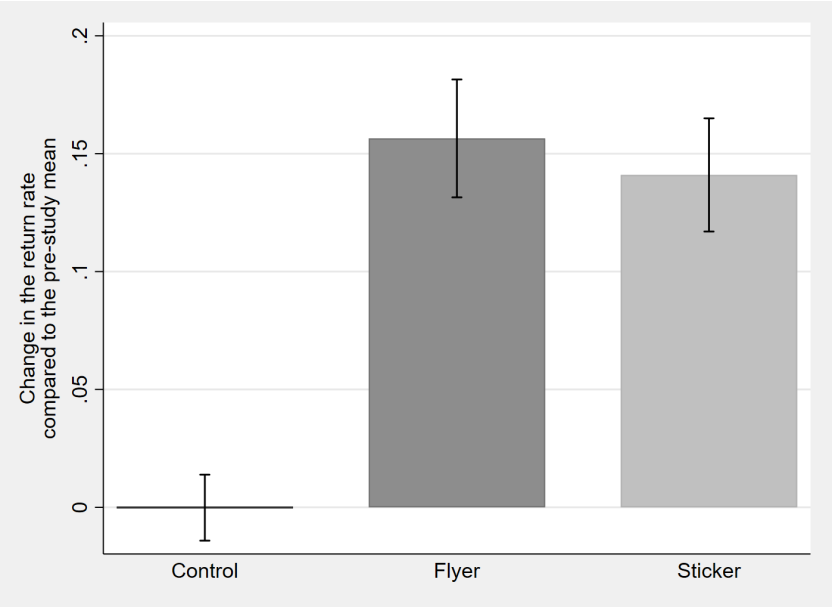
Notes: The table shows the average return rate across treatments in the pre-, post-, and intervention periods. Standard errors are reported in parentheses. The lower sample size in the post-intervention period is due to one customer who unsubscribed during the intervention.

While there are no significant differences across treatments before the intervention, we observe higher return rates in the flyer and sticker treatments than in the control group during the intervention period (p=0.000 flyer vs. control, p=0.000 sticker vs. control, two-sided t-test for unrelated samples). This also applies for the post-intervention period (p=0.000 flyer vs. control, p=0.003 sticker vs. control, two-sided t-test for unrelated samples). Figure 2.4 displays the mean differences in return rates between the intervention and pre-intervention periods. The graph shows that the return rate in the control group does not change over the periods, whereas we observe a sharp increase in the flyer and sticker reminder conditions. The standard error bars indicate highly significant differences between both experimental conditions and the control group, but not between the flyer and sticker treatments.

To analyse this effect in a more sophisticated way, we use the following difference-in-difference regression model:

$$\begin{aligned} y_{i,t} = \beta_0 + \beta_1 Flyer_i + \beta_2 Sticker_i + \beta_3 Period_t \\ + \beta_4 Flyer_i * Period_t + \beta_5 Sticker_i * Period_t + \epsilon_{i,t}, \end{aligned} \tag{2.1}$$

where $y_{i,t}$ is the return rate of customer i in period t. We consider two treatment dummy variables: the $Flyer_i$ dummy and the $Sticker_i$ dummy. Both treatment dummies are



Notes: Each bar indicates the change of the mean return rates during the intervention compared to the pre-intervention period. The error bars represent the mean ± the standard error of the mean.

FIGURE 2.4: Mean differences in return rates

0 for the control group and take on the value 1 if the customer is assigned to the flyer or the sticker treatment, respectively. We also include the common time effect $Period_t$, which is 1 for the intervention period and 0 for the pre-intervention period. Our main coefficients of interest are the interaction terms between the period dummy and the treatment dummies. These interaction terms indicate the differences in the pre- and intervention period return rates between the reminder treatments and the control group. In all model specifications, standard errors are clustered on the customer level.

Table 2.3 presents the estimated coefficients of Model 2.1. In line with the descriptive statistics, Specification 1 confirms the large and significant reminder effect. More specifically, the coefficients of the interaction terms between the period dummy and the flyer or sticker dummy are of similar magnitude and highly significant. Compared to the control group, the differential change over periods is 16 percentage points for the flyer group and 14 percentage points for the sticker group. This is equal to a relative increase in the return rates of 83% in the flyer treatment and 78% in the sticker treatment. The results suggest that the sticker and flyer reminders are similarly effective in promoting returning behavior. In fact, we cannot reject the null hypothesis that the flyer and the sticker treatments have the same impact on customers' return rates (p=0.801, Wald test).

In Specification 2, we additionally control for the number of returned food baskets per customer i in period t. As customers usually return all packaging materials together, it is not surprising that we observe a positive and significant association between returned food baskets and returned plastic bags. Importantly, the inclusion of this variable does

TABLE 2.3: Difference-in-difference estimation: Return rate per customer

	1	2
Flyer	0.031	0.032
	(0.035)	(0.035)
Sticker	0.037	0.040
	(0.037)	(0.037)
Period	-0.000	-0.068**
	(0.014)	(0.026)
Flyer x Period	0.157***	0.161***
	(0.025)	(0.025)
Sticker x Period	0.141***	0.154***
	(0.024)	(0.025)
Baskets returned		0.029***
		(0.010)
Observations	574	574
N customers	287	287
R^2	0.086	0.106

Notes: Specifications 1 and 2 present results of a difference-in-difference regression. Robust standard errors clustered on the individual level are in parentheses. The dependent variable is the plastic bag return rate per customer. *Flyer* and *Sticker* are dummy variables equal to 1 for customers in the flyer or sticker treatment, respectively, and 0 otherwise. The dummy variable *Period* is 1 for the intervention period and 0 for the pre-intervention period. Specification 2 further includes the number of returned food baskets. *, **, and *** document significance at the 10%, 5%, and 1% levels.

not alter the treatment effects.[29] To sum up, the flyer as well as the sticker reminders have a large impact on return rates. These reminder effects are similar to or higher than those detected in previous work (Werner et al. 1998, Osbaldiston and Schott 2012, Sussman et al. 2013, Altmann and Traxler 2014, Calzolari and Nardotto 2017).[30]

2.3.2 Reminder effects over time

Beyond the analysis of the entire intervention period, we are further interested in the development of the reminder effects over time. Figure 2.5 first provides descriptive evidence for the return rates over the delivery weeks. During the intervention period

[29] Considering the week of delivery in a difference-in-difference model with random effects for customers has no major bearing on the outcomes (see Table 2.8, Appendix B). As Table 2.9 in Appendix B further shows, the results stay robust when looking at the absolute number of plastic bags returned during the invention period in a Poisson regression model. Additional random effects regressions reveal that considering the week of return (in addition to the week of delivery) has no major impact. These results are available on request.

[30] In a related environmental context, Sussman et al. (2013), for example, find an increase of 64% in food waste composting behavior, while Werner et al. (1998) demonstrate an increase of 87% in polystyrene recycling.

(situated between the two vertical lines), customers strongly reacted to the reminders, with a peak return rate of 40% in the flyer treatment and 38% in the sticker treatment in week six.

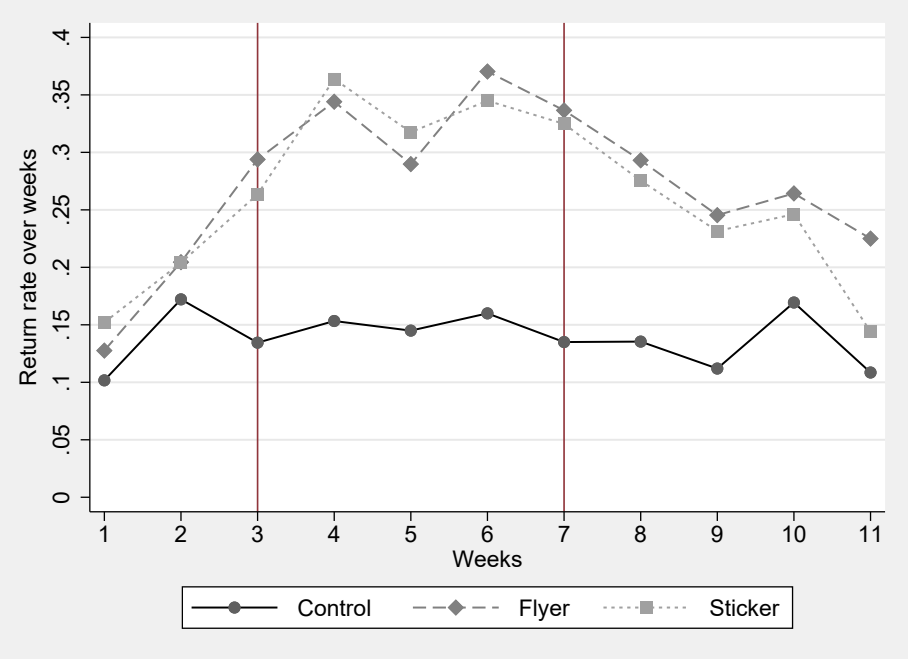


FIGURE 2.5: Return rates per treatment over weeks

To further study this time trend, we estimate the reminder effects during the intervention period with the following random effects model:

$$\begin{aligned} y_{i,t} = \beta_0 + \beta_1 Flyer_i + \beta_2 Sticker_i + \beta_3 Week_t \\ + \beta Flyer_{i,t} * Week_t + \beta Sticker_{i,t} * Week_t + \nu_i + \epsilon_{i,t}, \end{aligned} \tag{2.2}$$

where $y_{i,t}$ is the return rate of plastic bags that were delivered to customer i in week t. The predictors $Flyer_i$ and $Sticker_i$ are binary variables, showing customers' assignments to the flyer or sticker treatment. To capture the time trend in return behavior, we include the variable $Week_t$ (continuous, ranging from 3 to 7) and the corresponding interactions with the treatment dummies. The term ν_i indicates random, customer-specific deviations from the average, and $\epsilon_{i,t}$ is the random error term. Table 2.4 provides the estimates of Model 2.2.

The coefficient estimates for the flyer and sticker treatments are significant and of remarkable magnitude. These observed reminder effects do not differ from each other (Wald test p=0.953). Note that the treatment coefficients reflect the reminder effect in the first week of the intervention period. The interactions of the treatment dummies and the intervention weeks are positive but small in size and not significant. This confirms

TABLE 2.4: Random effects regression: Return rate per customer and delivery week

	1	2
Flyer	0.151**	0.117**
	(0.063)	(0.055)
Sticker	0.147***	0.117**
	(0.056)	(0.047)
Week	0.003	0.003
	(0.006)	(0.006)
Flyer x Week	0.007	0.007
	(0.010)	(0.010)
Sticker x Week	0.006	0.006
	(0.009)	(0.009)
Baskets returned		0.045**
		(0.019)
Return rate pre-intervention		0.800***
		(0.046)
Big box		0.052
		(0.038)
Meat box		0.047*
		(0.026)
Veggie box		0.059**
		(0.024)
FE Depot	No	Yes
Observations	1,390	1,390
N customers	287	287
sd (customers)	0.251	0.147
sd (residual)	0.205	0.204
R^2 overall	0.066	0.447

Notes: Specifications 1 and 2 present results of a random effects model with random effects for customers. Robust standard errors clustered on the individual level are in parentheses. The dependent variable is the return rate per customer and delivery week. *Flyer* and *Sticker* are dummy variables equal to 1 for customers in the flyer or sticker treatment, respectively, and 0 otherwise. The variable $Week$ is continuous, ranging from 3 to 7, and represents the delivery week in the intervention period. In Specification 2, the control variables include the number of baskets returned, the return rate in the pre-intervention period, dummy variables for the basket sizes (small boxes used as a reference) and the basket types (vegan boxes used as a reference), and fixed effects (FE) for depots. *, **, and *** document significance at the 10%, 5%, and 1% levels.

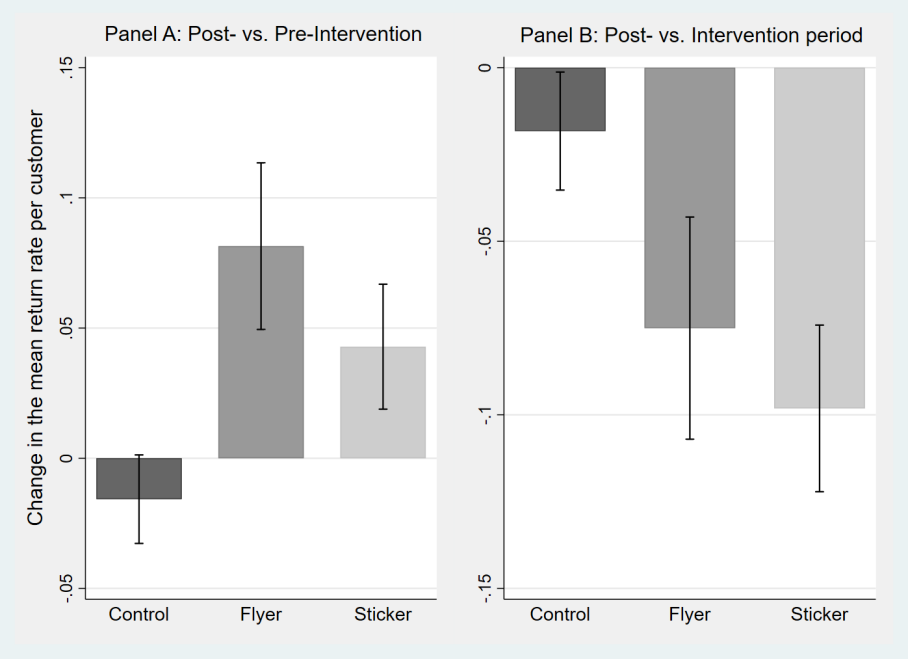
the temporal stability of the reminder effects during the intervention. As Specification 2 shows, both treatment coefficients slightly drop in size but stay significant at the 5% level when including depot-fixed effects and controlling for the pre-intervention return rate, the type, size, and number of returned food baskets. Again, we observe that the number of returned baskets has a positive and significant effect on the return rate of plastic bags.[31] Furthermore, customers with vegetarian food baskets ($Veggie\ box$) return

[31]The results stay robust when regressing the absolute number of plastic bags in a Poisson version of Model 2.2 (see Table 2.10 in Appendix B).

significantly more plastic bags for reuse than those with vegan baskets as a reference.

These time-related results suggest that reminders, regardless of their form, have a persistent positive effect on pro-environmental behavior, even when they are repeatedly applied. Customers do not seem to get used to the reminder information. This is congruent with earlier findings in environmental (Allcott and Rogers 2014) and non-environmental contexts (Kast et al. 2012, Apesteguia et al. 2013, Altmann and Traxler 2014, Calzolari and Nardotto 2017).[32]

We further investigate the post-intervention effects of our reminder treatments. Figure 2.6 delivers a graphical illustration of the mean differences in return rates during the post- vs. pre-intervention period and the post- vs. intervention period. Panel A shows a slight decrease in return behavior in the control group and a strong increase in both reminder treatments during the post-intervention period compared to the pre-intervention period. Panel B, on the other hand, illustrates the clear reduction of the post-intervention treatment effects, when compared to the intervention period.



Notes: Each bar indicates the change in mean return rates. The error bars represent the mean ± the standard error of the mean.

FIGURE 2.6: Mean differences in return rates in the post-intervention period

We also address these trends by applying a difference-in-difference model, similar to Model 2.1. Here, the $Period_t$ indicator takes the value of 1 for the post-intervention period and 0 for the pre-intervention period (Specifications 1 and 2) or for the intervention

[32] With respect to energy conservation, Allcott and Rogers (2014) show that consumers are very slow to habituate to reminders that include social feedback and still react to the messages two years after their introduction.

period (Specifications 3 and 4). Table 2.5 reports the estimates that support the initial impressions of Figure 2.6. Specifications 1 and 2 show that, in comparison to the pre-intervention period, the flyer and sticker reminders have a significant positive effect on return rates in the post-intervention period. This post-treatment effect provides some indication for habit formation. Nevertheless, Specifications 3 and 4 reveal that the post-intervention reminder effects are significantly lower than those observed during the intervention period. This is particularly the case for the sticker reminders, for which we observe a decrease of up to 8.9 percentage points (see Specification 4).[33]

TABLE 2.5: Post-intervention difference-in-difference regression

	Pre vs. Post		Intervention vs. Post	
	1	2	3	4
Flyer	0.035	0.037	0.189***	0.195***
	(0.035)	(0.035)	(0.038)	(0.037)
Sticker	0.041	0.044	0.179***	0.195***
	(0.037)	(0.037)	(0.038)	(0.038)
Post-intervention	-0.016	-0.047**	-0.018	0.015
	(0.016)	(0.021)	(0.013)	(0.016)
Flyer x Post-intervention	0.097***	0.096***	-0.057**	-0.063**
	(0.032)	(0.032)	(0.027)	(0.026)
Sticker x Post-intervention	0.059**	0.062**	-0.080***	-0.089***
	(0.024)	(0.025)	(0.020)	(0.020)
Baskets returned		0.026**		0.028***
		(0.010)		(0.008)
Observations	572	572	572	572
N customers	286	286	286	286
R^2	0.034	0.048	0.089	0.116

Notes: Specifications 1-4 present the results of a difference-in-difference regression with robust standard errors clustered on the individual level in parentheses. The dependent variable is the return rate per customer. The dummy variables *Flyer* and *Sticker* indicate the assignment of a customer to the flyer or sticker treatment, respectively. In Specifications 1 and 2, the dummy variable *Post-intervention* is 1 for the post-intervention period and 0 for the pre-intervention period. In Specifications 3 and 4, the dummy variable *Post-intervention* is 1 for the post-intervention period and 0 for the intervention period. Specifications 2 and 4 further include the number of returned food baskets. The lower sample size in the post-intervention period is due to one customer, who unsubscribed in the post-intervention period. *, **, and *** document significance at the 10%, 5%, and 1% levels.

Taken together, our findings suggest that reminder effects are at work beyond the intervention period, though to a smaller extent. This preliminarily applies to the flyer reminders. In fact, the flyer reminder effect detected in our study seems to endure longer than the reminder effects reported in previous work (Sussman and Gifford 2012, Allais et al. 2017, Calzolari and Nardotto 2017). However, once we stop enclosing reminders,

[33]Table 2.11 in Appendix B supports these results with separate estimates for each week of the post-intervention period. The flyer reminder seems to be slightly more persistent over time than the estimated sticker coefficients. However, except for the last week, we cannot reject the null hypothesis of identical effects at conventional significance levels.

the impact of both treatments decreases over time. This indicated decay is consistent with existing literature, showing that consumers are very slow in taking up new habits on the basis of reminders (Kast et al. 2012, Calzolari and Nardotto 2017) or on the basis of point-of-decision prompts (Sussman and Gifford 2012, Allais et al. 2017).

2.3.3 The impact of the reminders' proximity to action on return behavior

This section provides an in-depth analysis of whether a reminder's proximity to action has an effect on return behavior. In our setting, the sticker reminder functioned as an action-close reminder for plastic bags, where the reminder was directly attached, catching customers' attention when they decide to discard or return the plastic bag. For unmarked bags, on the other hand, the flyer and sticker reminders constituted conventional reminders. We thus investigate the question of whether a plastic bag with an action-close sticker reminder is more likely to be returned than any other, unmarked bag. We define the following logistic regression models:

$$ln(\frac{y_{b,i}}{1-y_{b,i}}) = \beta_0 + \beta_1 ActionCloseC_b + \delta_w + \nu_i + \epsilon_{b,i}, \tag{2.3a}$$

$$ln(\frac{y_{b,i}}{1-y_{b,i}}) = \beta_0 + \beta_1 ActionCloseF_b + \delta_w + \nu_i + \epsilon_{b,i}, \tag{2.3b}$$

$$ln(\frac{y_{b,i}}{1-y_{b,i}}) = \beta_0 + \beta_1 ActionCloseS_b + \delta_w + \nu_i + \epsilon_{b,i}, \tag{2.3c}$$

where $y_{b,i}$ is a dummy variable indicating whether the plastic bag b delivered to customer i has been returned or not. Because we want to differentiate the direct effect of the action-close sticker reminder, we confine our analysis to those bags with a sticker attached in the sticker treatment and use the unmarked bags of the control treatment (Model 2.3a), the flyer treatment (Model 2.3b), and the sticker treatment (Model 2.3c) as the comparison group. Accordingly, the indicators $ActionCloseC_b$, $ActionCloseF_b$, and $ActionCloseS_b$ are binary variables, taking the value 1 if plastic bag b had a sticker reminder attached (action-close reminder) and 0 for all unmarked bags from the control group, the flyer treatment, and the sticker treatment, respectively. In our analyses, we further include fixed effects for the delivery weeks (δ_w).[34] As before, the term ν_i indicates random effects at the customer level, and $\epsilon_{b,i}$ captures any other unmodeled effects.

Table 2.6 reports the estimated odds ratios of Models 2.3a, 2.3b, and 2.3c for the intervention period. The results show a strong positive and significant effect of the

[34] The outcomes are hardly affected if we do not control for the week of delivery. Results are available upon request.

action-close reminder on the probability that a plastic bag is returned. In Specification 1, the odds ratio of the action-close reminder indicates that the odds of returning are more than 14.7 times higher for bags with a sticker reminder attached than for bags in the control group. Specification 4 suggests that the odds of returning bags with an action-close reminder are 2.4 times higher than for bags in the flyer treatment. Adding the set of control variables in Specifications 2 and 5 further supports these findings. In Specifications 3 and 6, we additionally test the time effect by including the week variable as a continuous measure for the week of the intervention period (ranging from 3 to 7). As we can see, the interactions of the week variable and the dummy indicators $ActionCloseC_b$ or $ActionCloseF_b$ are insignificant, confirming the stability of the action-closeness effect over time.[35] The action-closeness effect is also supported by the estimates of Model 2.3c, where we compare the return rates of plastic bags with and without reminder stickers within the sticker treatment. Specification 9, for example, shows that the odds for returning a bag with a sticker attached are almost 5 times higher than the odds for returning an unmarked bag in the sticker treatment.[36]

[35] Step-by-step inclusion of control variables shows that our results are robust. Regressions available upon request.

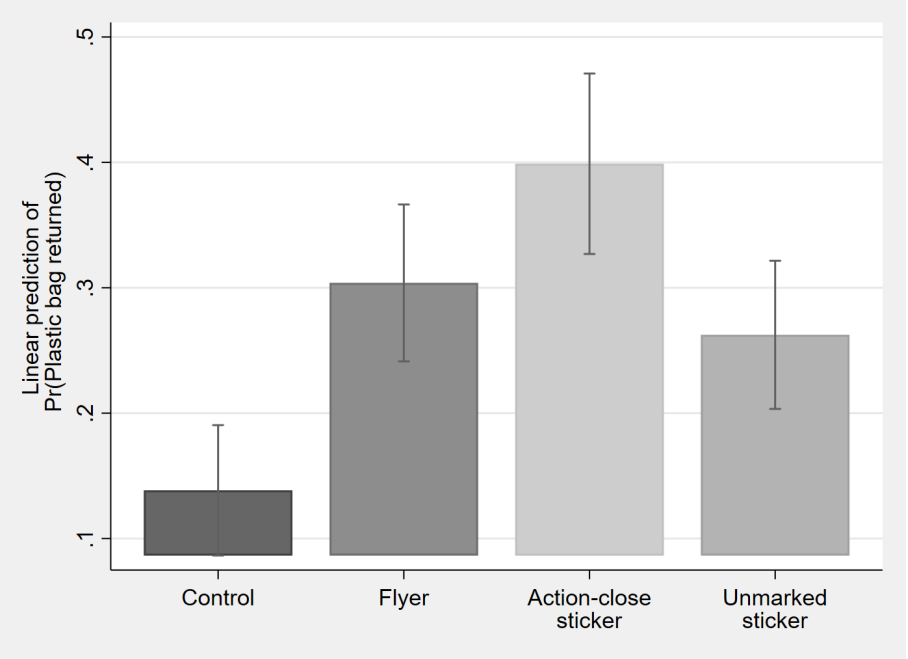
[36] We find similar and slightly stronger results when adding random effects for the delivery weeks in a mixed effects regression model (see Table 2.12 in Appendix B).

TABLE 2.6: Random effects logit regression: Odds that a plastic bag is returned

	Model 2.3a			Model 2.3b			Model 2.3c		
	1	2	3	4	5	6	7	8	9
Action-close	14.670***	12.023***	19.381***	2.410**	2.240***	5.103**	3.191***	3.179***	4.966***
	(7.090)	(3.989)	(13.650)	(0.875)	(0.595)	(3.259)	(0.467)	(0.461)	(2.704)
Baskets returned		1.058	1.057		1.092*	1.092*		1.003	1.003
		(0.057)	(0.056)		(0.053)	(0.053)		(0.053)	(0.052)
Return rate pre-intervention		188.537***	173.836***		103.817***	101.155***		280.545***	253.220***
		(112.295)	(103.139)		(61.579)	(59.688)		(173.210)	(154.014)
Big box		1.120	1.122		1.802*	1.810*		1.095	1.088
		(0.663)	(0.655)		(0.607)	(0.608)		(0.573)	(0.561)
Meat box		1.901	1.841		2.144**	2.099**		2.916**	2.830**
		(0.856)	(0.817)		(0.800)	(0.779)		(1.271)	(1.207)
Veggie box		1.399	1.411		1.961*	1.975*		1.836	1.866
		(0.539)	(0.536)		(0.687)	(0.688)		(0.716)	(0.719)
Week			1.097			1.164**			1.087
			(0.080)			(0.077)			(0.066)
Action-close x Week			0.895			0.844			0.903
			(0.109)			(0.097)			(0.085)
FE Depot	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
FE Delivery week	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No
Observations	2,081	2,076	2,076	2,108	2,108	2,108	1,920	1,892	1,892
N customers	143	142	142	144	144	144	75	74	74

Notes: Specifications 1–9 present the results of a logistic regression with random effects for customers. Robust standard errors clustered on the individual level are in parentheses. Estimates are presented in odds ratios. The dependent variable is a dummy variable indicating whether the plastic bag was returned or not. The dummy variable *Action-close* is 1 if the plastic bag had a sticker reminder attached and 0 for unmarked bags in the control group (Specifications 1–3), the flyer treatment (Specifications 4–6), and the sticker treatment (Specifications 7–9). In Specifications 2, 5, and 8, the control variables include the number of baskets returned, the return rate in the pre-intervention period, and dummy variables for the basket sizes (small boxes as a reference) and the basket types (vegan boxes as a reference). We also use fixed effects (FE) for the depots. In Specifications 1, 2, 4, 5, 7, and 8 we include fixed effects for the delivery week. Specifications 3, 6, and 9 include the week variable as a continuous measure for the weeks of the intervention period (ranging from 3–7). *, **, and *** document significance at the 10%, 5%, and 1% levels.

To illustrate the effect sizes, Figure 2.7 provides the predicted probabilities for returning a plastic bag with and without a sticker reminder in Specifications 1, 4, and 7. Whereas the probability for returning a bag with a sticker reminder attached is about 41%, the probability for returning an unmarked bag is 14% in the control, 29% in the flyer, and 26% in the sticker treatment. Using an action-close instead of a conventional reminder therefore increases the probability that a plastic bag is returned for reuse by 12–15 percentage points or, on a relative basis, by 41–58%. As suggested in prior studies (e.g., Sussman and Gifford 2012, Shearer et al. 2017), we also find a strong effect of action-closeness in comparison to the control group, where the difference in probability is approximately 26 percentage points or 200%.



Notes: Predicted probabilities that a plastic bag is returned, based on the logistic regression models 2.3a, 2.3b, and 2.3c. The error bars report the 95% confidence intervals.

FIGURE 2.7: Action-closeness effect

While we so far focused on the effect of an action-close sticker reminder on return behavior, we now further investigate the unmarked bags that were returned within the sticker treatment. The main question here is, to what extent the total impact of the sticker reminders was also driven by a spillover effect on unmarked plastic bags. We therefore compare the probabilities for returning plastic bags without a sticker across the experimental treatments and dismiss plastic bags with action-close stickers attached from the following analyses. Table 2.7 presents the estimates of the following logistic regression model, including random effects for customers:

$$ln(\frac{y_{ns,i}}{1-y_{ns,i}}) = \beta_0 + \beta_1 Flyer_i + \beta_2 Sticker_i + \nu_i + \epsilon_{ns,i}. \tag{2.4}$$

The dummy variable $y_{ns,i}$ indicates whether the unmarked plastic bag ns delivered to customer i has been returned or not. The indicators $Flyer_i$ and $Sticker_i$ are binary variables indicating the assignment of customer i to the flyer or sticker treatment, respectively. The term ν_i represents customer-specific random effects, and $\epsilon_{ns,i}$ is the random error term. We additionally use fixed effects to control for the delivery weeks.

Table 2.7 displays the estimated odds ratios of Model 2.4. As we can see in Specification 1, both reminder treatments have a positive and highly significant effect on the return probability of unmarked plastic bags compared to the control group. Focusing on the sticker treatment, we observe that the odds of returning a bag were 3.6 times higher for unmarked plastic bags in the sticker treatment than for bags in the control group. This indicates that the positive effect of the sticker treatment is not solely driven by bags with action-close stickers attached but also by spillovers of the sticker reminders to unmarked plastic bags. However, according to the Wald test, the sticker and flyer treatments seem to be similarly effective with respect to unmarked bags (p=0.293). We can also confirm that these results are robust, once we include our set of controls, and persistent over time (see Specifications 2 and 3).

Taken together, our findings suggest that a reminder's proximity to the decision situation is a crucial feature for its effectiveness. In our setting, we find that plastic bags with action-close sticker reminders attached are significantly more likely to be returned for reuse than unmarked bags in either of the treatment groups. This is an important and novel result, as previous studies are mainly limited to action-close prompts for encouraging pro-environmental behavior and do not compare the effects to conventional reminders (Austin et al. 1993, Houghton 1993, Sussman and Gifford 2012, Sussman et al. 2013, Shearer et al. 2017). Within the sticker treatment, we could further show that action-close reminders also increase the probability of returning unmarked bags for reuse. This leads to a similar improvement as the conventional flyer reminder. Such spillover effects were mentioned in previous work (e.g., Rea et al. 1987)[37] but have received minor attention so far. Overall, we conclude that action-close reminders have the potential to lever the benefits of conventional reminders on pro-environmental behavior, even when they are not applied to every single decision point.

[37] Rea et al. (1987) show that light usage in private offices is significantly reduced when reminder stickers are attached to light switch plates. This effect holds for offices with and without stickers.

TABLE 2.7: Random effects logit regression: Odds that an unmarked bag is returned

	1	2	3
Flyer	5.634***	4.692***	3.388**
	(2.582)	(1.394)	(1.899)
Sticker	3.661***	2.953***	3.063**
	(1.702)	(0.852)	(1.693)
Baskets returned		1.104**	1.103**
		(0.051)	(0.051)
Baseline return rate		342.328***	325.474***
		(158.825)	(150.076)
Big box		2.078**	2.084**
		(0.747)	(0.742)
Meat box		2.117**	2.049**
		(0.699)	(0.672)
Veggie box		1.892**	1.911**
		(0.577)	(0.579)
Week			1.099
			(0.081)
Flyer x Week			1.063
			(0.106)
Sticker x Week			0.993
			(0.095)
FE Depot	No	Yes	Yes
FE Delivery week	Yes	Yes	No
Observations	4,984	4,923	4,923
N customers	212	209	209

Notes: Specifications 1–3 present the results of a logistic regression with random effects for customers. Robust standard errors clustered on the individual level are in parentheses. Estimates are presented in odds ratios. The dependent variable is a dummy variable indicating whether a plastic bag without a sticker was returned or not. The dummy variables *Flyer* and *Sticker* indicate the assignment of a customer to the flyer or sticker treatment, respectively. In Specifications 2 and 3, the control variables include the return rate in the pre-intervention period, the number of baskets returned, and dummy variables for basket sizes (small boxes used as a reference) and basket types (vegan boxes used as a reference). We also include fixed effects (FE) for depots. In Specifications 1 and 2, we consider fixed effects for the delivery weeks, whereas Specification 3 includes the week variable as a continuous measure for the weeks of the intervention period (ranging from 3-7). *, **, and *** document significance at the 10%, 5%, and 1% levels.

2.4 Discussion

In this paper, we investigate the effect of reminders on pro-environmental behavior as measured by food box customers' propensity to return plastic bags for reuse. Our results show that weekly flyer and sticker reminders increase the return rates of plastic bags by up to 83% relative to a control treatment where no reminders were present. Both reminder treatments are similarly effective, indicating that the form of the reminder plays a subordinate role in promoting behavioral change (see Reekie and Devlin 1998). Interestingly, the reminder effects unfold from the beginning of the intervention, persist

over the entire intervention period, and last, to some degree, even beyond the intervention. Our study thus underlines the benefits of using simple reminders to enhance sustainable behavior.

Reusing plastics reflects an investment activity that is individually costly, while future environmental benefits are delayed and collectivized. In line with existing literature on behavior in investment tasks, our interpretation of the reminder effect is that reminders reduce customers' focus on the current salient costs of their behavior and increase the processing of future-related benefits (Borgstede and Andersson 2010, Karlan et al. 2016, Calzolari and Nardotto 2017). We complement this literature in a new setting, showing that reminders may help consumers to align their intentions with their actions regarding plastic waste.

We further contrast the return rate of plastic bags that were marked with an action-close sticker reminder and unmarked plastic bags in either of the treatment groups. We find that an action-close presentation has a strong effect on return rates for marked plastic bags. The probability of returning a bag with a sticker reminder directly attached is about 41%, while for unmarked bags it is about 29% in the flyer treatment, 26% in the sticker treatment, and 14% in the control group. This result shows that reminders are most effective when they bring desired behavior to the top of the mind in the decision situation and at the time when action is being taken. It thereby supports limited attention as an explanation for reminder effects. The fact that sticker reminders are also strongly effective for unmarked bags (i.e, when they are distant to the decision to reuse) confirms previous indications for the spillover effects of point-of-decision prompts (Rea et al. 1987).

Our results are in line with existing reminder interventions in other studies on pro-environmental behavior (Austin et al. 1993, Houghton 1993, Sussman and Gifford 2012, Sussman et al. 2013, Shearer et al. 2017). In a similar experiment, the findings of Austin et al. (1993) indicate that recycling behavior can be effectively encouraged if visual prompts and recycling containers are in close proximity to a recycling decision. Our study adds to this literature by using individual-level data and a design in which only the reminder's proximity to action is manipulated. We can thus provide compelling evidence for the action-closeness effect of reminders.

The reminder effects detected in our study are also economically significant. In our setting, a weekly flyer reminder leads to approximately 300 additional plastic bags returned per week. Such a behavioral change would imply that around 16,000 additional plastic bags are returned by the customers of the food box provider over the course of a year. With respect to the sticker intervention, marking every plastic bag with a sticker reminder would result in a significant reduction of plastic waste with approximately

25,000 additional plastic bags being reused instead of discarded per year. This, of course, assumes that the magnitude of the effect will persist in the long term.

Some limitations inherent to our setting raise open questions and provide opportunities for future research. While the sticker and flyer reminders were identical with respect to size and layout, participants may have perceived plastic bags with a sticker reminder attached as different, perhaps more valuable or important than unmarked bags. Neither of the reminders, however, extended the usage or purpose of the plastic bags or was inherently valuable for the customers. Nevertheless, it would be interesting to test recipients' perceived valuation of marked and unmarked plastic bags with regard to product reuse. Furthermore, the particular subject pool of our study may limit the potential to generalize our findings. Customers of regionally produced, organic food boxes presumably have an above-average awareness of environmental issues. This possibly reinforced the reminder effects in our analyses.[38] It would thus be interesting to re-examine the effect of action-close and conventional reminders in alternative settings and with different samples.

Our findings also indicate avenues for future research with relevance for policy makers. Several studies ask for research on the long-term effects of behavioral interventions (e.g., Steg and Vlek 2009, Croson and Treich 2014). While we show that customers continue to reuse plastic bags beyond the intervention period, return rates gradually decline after the end of the intervention. For policy design, this raises the question of at what intervals reminders should be presented to trigger long-term behavioral change. A second research area may be to investigate whether improved pro-environmental behavior in one area (e.g., reusing plastic bags) has the power to spill over into other environmental decisions. Daneshvary et al. (2016), for example, find that those who take part in curbside recycling are more likely to also take part in textile recycling schemes later on. Identifying conditions for and quantifying such spillover effects constitutes a promising field of future research (see Dolan and Galizzi 2015). Lastly, it would be interesting to understand the potential interaction effects of reminders with other motivational factors, such as commitment contracts (e.g., Can et al. 2003), financial incentives (e.g., Volpp et al. 2009), or information (e.g., Apesteguia et al. 2013, Altmann and Traxler 2014, Raifman et al. 2014) on behavioral change. Such interactions may exploit the benefits of reminders in addressing consumers' limited attention and possibly support pro-environmental behavior in the long term.

[38] Schultz (2014) shows that recycling prompts work most effectively for individuals with favorable attitudes toward recycling. According to Barr (2002), gender, family status, income, level of education, and political orientation affect recycling attitudes and behavior.

References

Acland, D. and Levy, M. R. (2015). Naiveté, projection bias, and habit formation in gym attendance. *Management Science*, 61(1):146–160.

Allais, O., Bazoche, P., and Teyssier, S. (2017). Getting more people on the stairs: The impact of point-of-decision prompts. *Social Science & Medicine*, 192:18–27.

Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *The American Economic Review*, 104(10):3003–3037.

Altmann, S. and Traxler, C. (2014). Nudges at the dentist. *European Economic Review*, 72:19–38.

Apesteguia, J., Funk, P., and Iriberri, N. (2013). Promoting rule compliance in daily-life: Evidence from a randomized field experiment in the public libraries of barcelona. *European Economic Review*, 64:266–284.

Attari, S. Z. (2014). Perceptions of water use. *Proceedings of the National Academy of Sciences of the United States of America*, 111(14):5129–5134.

Attari, S. Z., DeKay, M. L., Davidson, C. I., and Bruine de Bruin, W. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences of the United States of America*, 107(37):16054–16059.

Austin, J., Hatfield, D. B., Grindle, A. C., and Bailey, J. S. (1993). Increasing recycling in office environments: The effects of specific, informative cues. *Journal of Applied Behavior Analysis*, 26(2):247–253.

Barr, S. (2002). *Household Waste in Social Perspective: Values, Attitudes, Situation and Behaviour*. Ashgate, Aldershot.

Bisinella, V., Albizzati, P. F., Astrup, T. F., and Damgaard, A., editors (2018). *Life Cycle Assessment of grocery carrier bags. Danish Environmental Protection Agency*. Danish Environmental Protection Agency. Miljoeprojekter, No. 1985, Copenhagen.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.

Borgstede, C. v. and Andersson, K. (2010). Environmental information—explanatory factors for information behavior. *Sustainability*, 2(9):2785–2798.

Calzolari, G. and Nardotto, M. (2017). Effective reminders. *Management Science*, 63(9):2915–2932.

Can, S., Macfarlane, T., and O'Brien, K. D. (2003). The use of postal reminders to reduce non-attendance at an orthodontic clinic: a randomised controlled trial. *British Dental Journal*, 195(4):199–201.

Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.

Croson, R. and Treich, N. (2014). Behavioral environmental economics: Promises and challenges. *Environmental and Resource Economics*, 58(3):335–351.

Dahlbo, H., Poliakova, V., Mylläri, V., Sahimaa, O., and Anderson, R. (2018). Recycling potential of post-consumer plastic packaging waste in finland. *Waste Management*, 71:52–61.

Daneshvary, N., Daneshvary, R., and Schwer, R. K. (2016). Solid-waste recycling behavior and support for curbside textile recycling. *Environment and Behavior*, 30(2):144–161.

Dolan, P. and Galizzi, M. M. (2015). Like ripples on a pond: Behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47:1–16.

Geyer, R., Jambeck, J. R., and Law, K. L. (2017). Production, use, and fate of all plastics ever made. *Science Advances*, 3(7):1–5.

Hopewell, J., Dvorak, R., and Kosior, E. (2009). Plastics recycling: Challenges and opportunities. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1526):2115–2126.

Houghton, S. (1993). Using verbal and visual prompts to control littering in high schools. *Educational Studies*, 19(3):247–254.

Jacobs, H. E. and Bailey, J. S. (1982). Evaluating participation in a residential recycling program. *Journal of Environmental Systems*, 12(2):141–152.

Jae, H., Delvecchio, D. S., and Cowles, D. (2008). Picture-text incongruency in print advertisements among low- and high-literacy consumers. *Journal of Consumer Affairs*, 42(3):439–451.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment under uncertainty*. Cambridge University Press, Cambridge.

Karlan, D., McConnell, M., Mullainathan, S., and Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. *Management Science*, 62(12):3393–3411.

Kast, F., Meier, S., and Pomeranz, D. (2012). *Under-Savers Anonymous: Evidence on Self-Help Groups and Peer Pressure as a Savings Commitment Device*. National Bureau of Economic Research, Cambridge, MA.

Kline, T. J. B. and Beitel, G. A. (2016). Assessment of push/pull door signs: A laboratory and field study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 36(4):684–699.

Krendl, K. A., Olson, B., and Burke, R. (1992). Preparing for the environmental decade: A field experiment on recycling behavior. *Journal of Applied Communication Research*, 20(1):19–36.

Luyben, P. D. (1980). Effects of informational prompts on energy conservation in college classrooms. *Journal of Applied Behavior Analysis*, 13(4):611–617.

MacArthur, E. (2017). Beyond plastic waste. *Science*, 358(6365):843.

Miafodzyeva, S. and Brandt, N. (2013). Recycling behaviour among householders: Synthesizing determinants via a meta-analysis. *Waste and Biomass Valorization*, 4(2):221–235.

Osbaldiston, R. and Schott, J. P. (2012). Environmental sustainability and behavioral science. *Environment and Behavior*, 44(2):257–299.

Raifman, J. R. G., Lanthorn, H. E., Rokicki, S., and Fink, G. (2014). The impact of text message reminders on adherence to antimalarial treatment in northern ghana: a randomized trial. *PloS One*, 9(10):1–10.

Rea, M. S., Dillon, R. F., and Levy, A. W. (1987). The effectiveness of light switch reminders in reducing light usage. *Lighting Research & Technology*, 19(3):81–85.

Reekie, D. and Devlin, H. (1998). Preventing failed appointments in general dental practice: A comparison of reminder methods. *British Dental Journal*, 185(9):472–474.

Roberts, N. J., Mohamed, Z., Wong, P.-S., Johnson, M., Loh, L.-C., and Partridge, M. R. (2009). The development and comprehensibility of a pictorial asthma action plan. *Patient Education and Counseling*, 74(1):12–18.

Russell, W. D., Dzewaltowski, D. A., and Ryan, G. J. (1999). The effectiveness of a point-of-decision prompt in deterring sedentary behavior. *American Journal of Health Promotion*, 13(5):257–259.

Schultz, P. W. (2014). Strategies for promoting proenvironmental behavior. *European Psychologist*, 19(2):107–117.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Shearer, L., Gatersleben, B., Morse, S., Smyth, M., and Hunt, S. (2017). A problem unstuck? evaluating the effectiveness of sticker prompts for encouraging household food waste recycling behaviour. *Waste Management*, 60:164–172.

Steg, L. and Vlek, C. (2009). Encouraging pro-environmental behaviour: An integrative review and research agenda. *Journal of Environmental Psychology*, 29(3):309–317.

Stein, R. S. (1992). Polymer recycling: Opportunities and limitations. *Proceedings of the National Academy of Sciences*, 89(3):835–838.

Sussman, R. and Gifford, R. (2012). Please turn off the lights: The effectiveness of visual prompts. *Applied Ergonomics*, 43(3):596–603.

Sussman, R., Greeno, M., Gifford, R., and Scannell, L. (2013). The effectiveness of models and prompts on waste diversion: A field experiment on composting by cafeteria patrons. *Journal of Applied Social Psychology*, 43(1):24–34.

Szilagyi, P. G. and Adams, W. G. (2012). Text messaging: A new tool for improving preventive services. *Journal of the American Medical Association*, 307(16):1748–1749.

Taubinsky, D. (2013). From intentions to actions: A model and experimental evidence of inattentive choice. Working Paper, Harvard University. Cambridge, MA.

Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. (2018). Overcoming salience bias: How real-time feedback fosters resource conservation. *Management Science*, 64(3):1458–1476.

United Nations Environment Programme [UNEP] (2018). *Single-use Plastics: A Roadmap for Sustainability.* United Nations Environment Programme. Retrieved from https://wedocs.unep.org/bitstream/handle/20.500.11822/25496/singleUsePlastic_sustainability.pdf?isAllowed=y&sequence=1.

Volpp, K. G., John, L. K., Troxel, A. B., Norton, L., Fassbender, J., and Loewenstein, G. (2008). Financial incentive-based approaches for weight loss: A randomized trial. *Journal of the American Medical Association*, 300(22):2631–2637.

Volpp, K. G., Troxel, A. B., Pauly, M. V., Glick, H. A., Puig, A., Asch, D. A., Galvin, R., Zhu, J., Wan, F., DeGuzman, J., Corbett, E., Weiner, J., and Audrain-McGovern, J. (2009). A randomized, controlled trial of financial incentives for smoking cessation. *The New England Journal of Medicine*, 360(7):699–709.

Werner, C. M., Cook, S., Colby, J., and Lim, H.-J. (2012). "Lights out" in university classrooms: Brief group discussion can change behavior. *Journal of Environmental Psychology*, 32(4):418–426.

Werner, C. M., Rhodes, M. U., and Partain, K. K. (1998). Designing effective instructional signs with schema theory: Case studies of polystyrene recycling. *Environment and Behavior*, 30(5):709–735.

Williams, M., Thyer, B. A., Bailey, J. S., and Harrison, D. F. (1989). Promoting safety belt use with traffic signs and prompters. *Journal of Applied Behavior Analysis*, 22(1):71–76.

Appendix A Additional figures

"Please return the plastic bags. They can be reused."

FIGURE 2.8: Reminder message

Notes: The labeling procedure worked as follows. We affixed a tag with multiple transparent ID labels to each of the baskets, before they were filled. From this tag, the ID labels could be peeled and quickly attached to the plastic bags of the corresponding customer when the vegetables were put inside.

FIGURE 2.9: Plastic bag labeling

Appendix B Robustness checks

TABLE 2.8: Difference-in-difference random effects regression: Return rate per week

	1	2
Flyer	0.033	0.035
	(0.033)	(0.034)
Sticker	0.043	0.046
	(0.035)	(0.035)
Intervention	0.010	0.011
	(0.014)	(0.014)
Flyer x Period	0.150***	0.149***
	(0.027)	(0.027)
Sticker x Period	0.135***	0.137***
	(0.025)	(0.025)
Baskets returned		0.039**
		(0.017)
sd (customers)	0.233	0.230
sd (residual)	0.207	0.207
R^2 overall	0.072	0.080
Observations	1,949	1,949
N customers	287	287

Notes: The table displays the results of a difference-in-difference regression with random effects for customers. Robust standard errors clustered on the individual level are in parentheses. The dependent variable is the plastic bag return rate per customer per delivery week. *Flyer* (*Sticker*) is a dummy variable equal to 1 for customers in the flyer (sticker) treatment and 0 otherwise. The dummy variable *Period* is 1 for the intervention period and 0 for the pre-intervention period. Specification 2 further includes the number of returned food baskets. *, **, and *** document significance at the 10%, 5%, and 1% levels.

TABLE 2.9: Poisson regression: Plastic bags returned during the intervention

	1	2
Flyer	0.832***	0.818***
	(0.182)	(0.179)
Sticker	0.788***	0.807***
	(0.185)	(0.182)
Plastic bags delivered	0.036**	0.054*
	(0.017)	(0.029)
Baskets returned		0.113***
		(0.042)
Big box		0.084
		(0.213)
Meat box		0.378
		(0.244)
Veggie box		0.383**
		(0.178)
FE Depot	No	Yes
Observations	287	287

Notes: The table reports the results of a Poisson regression with robust standard in parentheses. The dependent variable is the number of plastic bags returned per customer during the intervention period. *Flyer* (*Sticker*) is a dummy variable equal to 1 for customers in the flyer (sticker) treatment and 0 otherwise. *Plastic bags delivered* indicates the number of plastic bags a customer received during the intervention period. Specification 2 further includes control variables for the number of baskets returned, dummy variables for the basket sizes (small boxes as a reference) and the basket types (vegan boxes as a reference), and fixed effects (FE) for depots. *, **, and *** document significance at the 10%, 5%, and 1% levels.

TABLE 2.10: Random effects Poisson regression: Plastic bags returned over time

	1	2
Flyer	0.864***	1.126***
	(0.314)	(0.287)
Sticker	0.809***	1.033***
	(0.313)	(0.288)
Delivery week	0.043	0.020
	(0.038)	(0.038)
Flyer x Delivery week	-0.002	-0.004
	(0.046)	(0.046)
Sticker x Delivery week	-0.001	-0.005
	(0.046)	(0.046)
Pastic bags delivered		0.234***
		(0.022)
Baseline returns		0.347***
		(0.031)
Baskets returned		0.164*
		(0.088)
Big box		0.317
		(0.242)
Meat box		0.556***
		(0.197)
Veggie box		0.626***
		(0.178)
$\ln(\alpha)$	0.639***	-0.105
	(0.106)	(0.135)
FE Depot	No	Yes
Observations	1,390	1,390
N customers	287	287

Notes: The table reports the results of a Poisson regression with random effects for customers. Robust standard errors clustered on the individual level are in parentheses. The dependent variable is the number of plastic bags returned per customer per delivery week. *Flyer* (*Sticker*) is a dummy variable equal to 1 for customers in the flyer (sticker) treatment and 0 otherwise. *Plastic bags delivered* indicates the number of plastic bags a customer received during the intervention period. Specification 2 includes control variables for the number of baskets returned, dummy variables for the basket sizes (small boxes as a reference) and the basket types (vegan boxes as a reference), and fixed effects (FE) for depots. *, **, and *** document significance at the 10%, 5%, and 1% levels.

TABLE 2.11: Regressions per post-intervention week

	Week 8	Week 9	Week 10	Week 11
Flyer	0.158***	0.133***	0.095*	0.116***
	(0.045)	(0.040)	(0.051)	(0.043)
Sticker	0.140***	0.120***	0.077	0.035
	(0.041)	(0.038)	(0.051)	(0.039)
Observations	278	282	262	262
R^2	0.052	0.045	0.014	0.030

Notes: The table presents separate OLS regressions of the return rate per client for each delivery week of the post-intervention period. *Flyer* and *Sticker* are dummy variables equal to 1 for customers in the flyer or sticker treatment, respectively, and 0 otherwise. We have missing observations for some customers in the individual delivery weeks due to holidays. *, **, and *** document significance at the 10%, 5%, and 1% levels.

TABLE 2.12: Multilevel logistic regression: Odds that a plastic bag is returned

	Model 2.3a		Model 2.3b		Model 2.3c	
	1	2	3	4	5	6
Action-close	19.838***	39.797***	2.684**	6.992**	4.012***	8.194***
	(10.832)	(36.441)	(1.102)	(5.317)	(0.741)	(5.404)
Week		1.125		1.187**		1.105
		(0.100)		(0.080)		(0.093)
Action-close x Week		0.871		0.826		0.868
		(0.126)		(0.105)		(0.109)
var (customers)	7.022	7.009	4.542	4.517	6.387	6.403
var (weeks)	1.254	1.220	1.144	1.060	1.892	1.890
Observations	2,081	2,081	2,108	2,108	1,920	1,920
N customers	143	143	144	144	75	75

Notes: Specifications 1–6 present the results of a logistic regression with random effects for customers and for delivery weeks (three levels). Robust standard errors are clustered on the customer level and are reported in parentheses. Estimates are presented in odds ratios. The dependent variable is a dummy variable indicating whether the plastic bag was returned or not. The dummy variable *Action-close* is 1 if the plastic bag had a sticker reminder attached and 0 for unmarked bags in the control group (Specifications 1–2), the flyer treatment (Specifications 3–4), and the sticker treatment (Specifications 5–6). Specifications 2, 4, and 6, include the week variable as a continuous measure for the weeks of the intervention period (ranging from 3–7). *, **, and *** document significance at the 10%, 5%, and 1% levels.

Essay 3: Motivating volunteers to engage more actively in charity: A field experiment

Angela Steffen, Zita Spillmann [*]

Abstract

Charitable organizations often face shortages in volunteer staff. In a randomized controlled trial, we test whether simple performance information can motivate volunteers of a German aid organization to increase their prosocial engagement. Newsletters informing volunteers about the average required service hours per volunteer should especially encourage less-active members and thus support a more balanced distribution of the organization's workload. Our results, however, show that this intervention has no significant effect. On the contrary, less-active volunteers tend to decrease their number of working hours, and more-active volunteers tend to increase their engagement, reinforcing inequality of the workload distribution. Additional analyses support the idea that organizational commitment drives the effectiveness of internal communication with volunteers.

Keywords: volunteer motivation, charitable organization, communication, performance information, prosocial behavior

[*]Angela Steffen: angela.steffen@iop.unibe.ch, Institute for Organization and Human Resource Management, University of Bern. Zita Spillmann: zita.spillmann@outlook.com. We are grateful to the Malteser Hilfsdienst e.V. Berlin for enabling this project. In particular, we thank Gereon Schomacher and Sandro Jasker for their cooperation and support. We are also grateful to Andrea Essl, Frauke von Bieberstein and the participants of the CUSO Summer School 2017 for the valuable discussions and feedback. All errors are ours.

3.1 Introduction

Volunteerism is an indicator of a country's welfare, providing social as well as economic benefits (BFS 2017a). It can be defined as "long-term, planned, prosocial behavior that benefits strangers and occurs within an organizational setting" (Penner 2002, p. 448). Due to its social implications, active participation in voluntary work has been acknowledged as an expression of the cohesion of a society and as an important generator of social capital (Putnam 1995, Coleman 2000).

In Germany, volunteerism is a widespread and culturally anchored part of civil society. In 2014, 43.6% of the population over the age of 14 years was actively engaged in institutionalized or informal voluntary activity (Simonson et al. 2017). Although the number of people involved in volunteering has increased during the past 15 years, volunteers devote less and less time to their engagement (BFS 2017b, Simonson et al. 2017). Volunteer organizations are therefore increasingly challenged to motivate their members to actively engage in voluntary services. This is the case for Malteser Hilfsdienst e.V. in Berlin (Malteser Berlin), a German aid organization that provides emergency prevention, hospice work, and medical assistance, among other things. On the one hand, Malteser Berlin has difficulties motivating less-engaged members to register for the organization's first-aid and care services (see Section 3.3.1). These services are offered for public events and pre-contracted with external event providers. The organization must therefore ensure that all the assignments can be staffed with the required number of volunteers. On the other hand, management at Malteser Berlin is worried that highly engaged volunteers become overburdened, since they often fill in when personnel bottlenecks occur. Malteser Berlin therefore aims for a balanced distribution of its workload by encouraging increased participation from less-active volunteers and relief for more-active volunteers.

In this study, we introduce simple performance information about the required engagement per volunteer to address this issue. In a randomized controlled trial, two different types of email newsletters were sent to 305 volunteers in the emergency-prevention department of Malteser Berlin. These newsletters informed the members about the availability of new voluntary assignments that were to be staffed. The emails to the treatment group contained additional information on the number of service hours that each volunteer would have to perform, if the workload of the next 30 days was going to be be equally distributed among the volunteers. We expect that this kind of normative performance information supports prosocial behavior at Malteser Berlin. By motivating less-active volunteers, we hypothesize that the intervention 1) increases the number of voluntary service hours and 2) balances the members' workload in comparison to the control

group. We draw these hypotheses from the previous research relating to norm conformity, altruistic motives, and self-efficacy.

First, the provided performance information may activate social expectations, as it signals a desired state or a kind of social norm (Schwartz 1965, pp. 224-225). A considerable amount of research shows that the perception of what others believe to be appropriate conduct—so-called injunctive social norms—has a strong impact on behavior (Cialdini et al. 1990; see Cialdini and Trost 1998, Cialdini and Goldstein 2004 for extensive reviews on the power of conformity). Normative information also seems to play an important role in goal-setting, as individuals tend to adjust their self-set goals to their "belief of what is appropriate or desirable" (Latham and Locke 1991, p. 220). In the context of prosocial behavior, the impact of social standards is confirmed by theoretical as well as empirical contributions (Vesterlund 2003, Frey and Meier 2004, Bénabou and Tirole 2006, Shang and Croson 2009). According to socio-psychological evidence, conformity to social norms is particularly strong when all group members work toward a common goal (Deutsch and Gerard 1955, Allen 1965) or when people in the group are similar or are friends (Lott and Lott 1961, Abrams et al. 1990). Both conditions apply to our setting. Social comparison research further shows that individuals especially rely on social standards in situations that are ambiguous (Buunk and Mussweiler 2001, Suls et al. 2016). Since Malteser Berlin has no commitment contracts, goals, or guidelines that specify the expected effort of their members, we believe that our intervention provides an important behavioral benchmark. In line with the results of Chen et al. (2010), this should mainly increase the motivation of less-active volunteers. Volunteers with above-average engagement, on the other hand, already comply with the desired behavior and may thus be less affected.[39]

The second idea that underlies our hypotheses is the volunteers' motivation to help others. While the reasons for voluntary engagements at Malteser Berlin are diverse, we identified "helping others" as a strong shared value and a common purpose among its members. Existing literature confirms that concern for others is a primary motivator for volunteering and an important personality characteristic of volunteers (e.g., Allen and Rushton 2016, Anderson and Moore 2016). Our intervention aims to transfer this altruistic concern into action by suggesting that the organization itself, or other members, require help. Various interviews with the organization's management revealed that strong social ties exist between volunteers. We therefore expect that the performance information appeals to the members' sense of community and altruistic concern for friends, reinforcing their

[39] In a field experiment on public-good contributions, Chen et al. (2010) argue that the conformity effect is attenuated for above-median performers, because individuals perceive contributing as a socially desirable course of action. This triggers competitive preferences (i.e., more effort is better), explaining why the contributions of the above-median group stay high when the (lower) performance average is revealed. For the below-median group, competitiveness and conformity both support increasing contributions.

engagement. This idea is supported by existing evidence that highlights the importance of presenting the target group as needy in order to increase volunteer participation (Fisher and Ackerman 1998). With respect to less-active members, information about the required contribution per volunteer may also stimulate feelings of guilt (e.g., Bolton and Ockenfels 2000) or a sense of duty due to inequity aversion (e.g., Fehr and Schmidt 1999).[40]

Finally, we expect our intervention to improve the workload balance at Malteser Berlin by enhancing members' self-efficacy.[41] Even though self-efficacy is mostly considered as a moderating variable, one can assume a direct causal relationship between self-efficacy and the intention to perform specific activities (Bandura 1977, Locke et al. 1990). At Malteser Berlin, we discovered that some volunteers have difficulties reconnecting to the Malteser community once they stopped regularly registering for voluntary assignments. Our intervention aims to increase the perceived self-efficacy of inactive volunteers by breaking down the workload into a specific, manageable amount of time and by demonstrating that even small contributions are meaningful and appreciated. Prior studies confirm that self-efficacy interventions can significantly increase peoples' willingness to volunteer (Eden and Kinnar 1991, Lindenmeier 2008, Martinez and McMullin 2016).

In contrast to these theoretical considerations and our hypotheses, we find no significant effect of our intervention on the engagement of less-active volunteers. More-engaged members seem to be even more reactive to the performance information, than those with a below-average commitment, and tend to increase their hours of service. Therefore, information about the required contribution per volunteer rather promotes inequality of the workload distribution at Malteser Berlin. While this effect is not significant, additional pretest–posttest analyses indicate that the newsletters in general have a significant negative impact on the workload balance.

The conclusions we draw from these findings and our experimental design contribute to existing literature in the following ways. First, our study complements prior research around social performance feedback and employee motivation (e.g., Blanes i Vidal and Nossol 2011, Delfgaauw et al. 2013), as we investigate the impact of work-related information in a non-commercial setting. This is of practical importance, as a growing number of studies question the transferability of paid-staff management practices to the volunteer sector (Machin and Paine 2008, Barnes and Sharpe 2009, Studer and

[40] Inequity aversion implies that people are willing to give up some material payoff to move in the direction of more equitable outcomes (Fehr and Schmidt 1999, p. 819). Empirical research shows that subjects exhibit a strong aversion against advantageous as well as disadvantageous inequality (e.g., Loewenstein et al. 1989).

[41] Self-efficacy refers to a person's expectation about being able to successfully carry out desired actions on his or her own (Bandura 1982, Stajkovic and Luthans 1998).

von Schnurbein 2013, p. 410). Second, academic research exploring interventions that stimulate the individual involvement of volunteers is still scarce (see Studer and von Schnurbein 2013, Graf 2015). Existing socio-psychological literature focuses on the motives (e.g., Cnaan and Goldberg-Glen 1991, Clary et al. 1998) and the individual dispositions or characteristics of volunteers (e.g., Penner and Finkelstein 1998, Bussell and Forbes 2002, Penner 2002, Carpenter and Myers 2010). In this study, in contrast, we test a simple and cost-effective intervention to increase volunteers' willingness to work. New findings of this sort can be applied in various voluntary settings, where financial resources are usually limited. Finally, we broaden the common objective of enhancing prosocial behavior by also considering distributional aspects. Since the total workload in our study is exogenously fixed (see Section 3.3.1), we focus on a situation where voluntary contributions are clear substitutes (see Warr 1982 and Roberts 1984 for models where public spending substitutes for private donations). Most empirical studies on charitable donations and public-good provisions assume a positive, that is, a complementary relationship between others' contributions and one's own (Keser and van Winden 2000, Fischbacher et al. 2001, Shang and Croson 2009).[42] The balancing of substitutable voluntary contributions has received minor attention so far.

The remainder of this paper is organized as follows. Beyond the theoretical foundations above, Section 3.2 provides additional insights into the literature around volunteer motivation. The field setting, experimental design, and data are described in Section 3.3. Section 3.4 presents the results, and the findings are discussed in Section 3.5.

3.2 Literature on volunteer motivation

Since volunteer activities, by definition, do not anticipate any direct consideration or monetary recompense, numerous papers from social-psychology and economics direct their attention to the motives for volunteering. These motives turn out to be highly diverse, entailing altruistic and egoistic components (Cnaan and Goldberg-Glen 1991, Clary et al. 1998).[43] The studies on why people volunteer are supplemented by literature around individual and organizational characteristics that influence volunteers' commitment and engagement. Carpenter and Myers (2010), for example, find that the decision to volunteer

[42] In Ziemek's (2006) model, volunteers regard their own donations and collective donations by others as substitutes, complements, or neither, depending on whether they follow altruistic, investment, or private consumption motivations.

[43] Volunteers may be driven by altruistic or humanitarian concerns for others (e.g., Clary and Miller 1986), look for new learning experiences (e.g., Gidron 1978), or strive for career-related benefits (e.g., Jenner 2016). Other studies suggest that people may engage in voluntary work to reduce their guilt over their superior situation and protect their ego (e.g., Frisch and Gerrard 1981). Voluntary work also has a social function by offering opportunities to be with one's friends or to engage in an activity viewed favourably by others (e.g., Rosenhan 1970).

is positively related to concerns for social reputation and altruism. Penner et al. (1995) and Penner and Finkelstein (1998) show that other-oriented empathy and helpfulness, two characteristics that determine a prosocial personality, are important determinants for the level of voluntary engagement. With regard to organizational factors, Cnaan and Cascio (1998) suggest that supervision and contact by mail or telephone is positively related to volunteers' satisfaction and tenure. Stirling et al. (2011) confirm the importance of communication by showing that publicly recognizing volunteers, for example through a newsletter, has a positive impact on the acquisition of new members.

With a closer focus on existing members, further studies demonstrate that volunteers who are more committed to the organization's concern or have a stronger identification with the organization's mission show higher levels of activity (Craig-Lees et al. 2008, Hustinx and Lammertyn 2016). Bennett and Barkensjo (2005) provide evidence that volunteers' organizational commitment is affected by internal marketing strategies, that is, the organization's internal communication, information sharing, and volunteer training. With respect to communication, Clary et al. (1994) find that matching the predominant motive for volunteering with a persuasive message that responds to that motive increases volunteers' willingness to work. Boezeman and Ellemers (2008, p. 1013) emphasize inducing "anticipated feelings of respect" to increase the attractiveness of a volunteer organization and to motivate new members to volunteer their services. Investigating the different communication policies of volunteer organizations, Lindenmeier (2008) shows that advertisement-induced emotional arousal, message framing, and manipulations of self-efficacy perceptions impact peoples' willingness to volunteer. Eden and Kinnar (1991) further demonstrate in a field experiment with candidates for special-forces service that enhancing self-efficacy through verbal persuasion and vicarious experience increases voluntary activity by raising workers' self-expectations. Overall, these studies suggest that human resource practices and, in particular, internal communication strategies crucially influence volunteers' engagement.

More specific evidence on motivating volunteers through performance-related information is scarce. With reference to social norms, Chen et al. (2010) examine content contribution in an online community and find that social information on average contribution levels motivates below-average performers to increases their contributions. Members with an above-average performance, on the other hand, show no significant reaction. In the context of a fundraising campaign, Frey and Meier (2004) show that information about average past donations has a significant impact positive on charitable giving. Shang and Croson (2009) further demonstrate that the most influential social information to maximize donations is contribution behavior drawn from the ninetieth to ninety-fifth percentile. Our experiment differs from these studies by investigating volunteering time (see Bénabou and Tirole 2006), instead of charitable donations, and by using

future-oriented information about desired contributions. This information resembles an injunctive norm of what others would approve rather than a descriptive norm about what others actually do (see Schultz et al. 2007).

3.3 Methodology

3.3.1 Field setting

This field experiment was conducted in cooperation with the charitable organization Malteser Berlin. More than 1,000 volunteers and more than 400 full-time employees work for Malteser Berlin in various services associated with charity. This study focuses on volunteers in the emergency-prevention department, whose main responsibility is nationwide civil protection in the case of a disaster. On a day-to-day basis, the department mainly provides first-aid and care services at public events, such as concerts, theatre performances, or sporting events. This ensures regular training of the volunteers' medical skills and generates financial revenues, which Malteser Berlin uses for other charitable projects. At the time of our study, the emergency-prevention department counted 375 volunteer members.

The head coordinator of the department contracts the service requests of event organizers (i.e., the customers) several months before the event date. Depending on the location and the size of the event, the head coordinator allocates accepted assignments to either one or all three divisions of the emergency-prevention department (i.e., Berlin-North, Berlin-South, and Berlin-West).[44] The staffing process then works as follows. Each volunteer is registered on an internal online platform called Hioplan. On Hioplan, volunteers can sign up or opt out for open assignments, via their personal login. The event coordinators of the three divisions regularly update the assignments and supervise the registration process. Since Malteser Berlin contracts the necessary resources upfront, the workload is fixed, and the divisional event coordinators are responsible for filling the events with the required number of volunteers.

[44] In our sample, 44% of the assignments were open to all three divisions, 6% to Berlin-North, 36% to Berlin-South, and 13% to Berlin-West. The divisional affiliation of the volunteers is shown in Table 3.1 in Section 3.3.3.

3.3.2 Experimental design and procedure

Using a between-subjects design, we sent two types of newsletters to all 305 volunteers in the emergency-prevention department who engage in care or first-aid services.[45] These volunteers were randomly split into the control and treatment group, using a stratified sampling method (see Bruhn and McKenzie 2009). Since there exist important differences in the number of available assignments between the geographic sections, we first split all participants into the divisions Berlin-North, -South, and -West. To ensure a more balanced distribution of prior activity levels, we further sorted these subsets by the volunteers' total number of logins to Hioplan. We used this variable as a proxy for the volunteers' previous engagement, as this information was not available at the beginning of our study.

The newsletters were newly created for this project and sent to the volunteers' private email addresses. All emails reminded volunteers about the availability of new assignments and suggested two specific events for which to sign up. The treatment group additionally received information on the number of service hours that each volunteer would have to perform, if the workload of the next 30 days was equally distributed. To determine this number, the duration of all announced assignments for the following 30 days was added up and divided by the 305 participants. This average was then rounded off to the next half hour. Figure 3.1 provides an example of a newsletter sent to the treatment group with the performance information in bold font. The original German newsletters can be found in Appendix A. The emails to the control group were the same, except for the performance information.

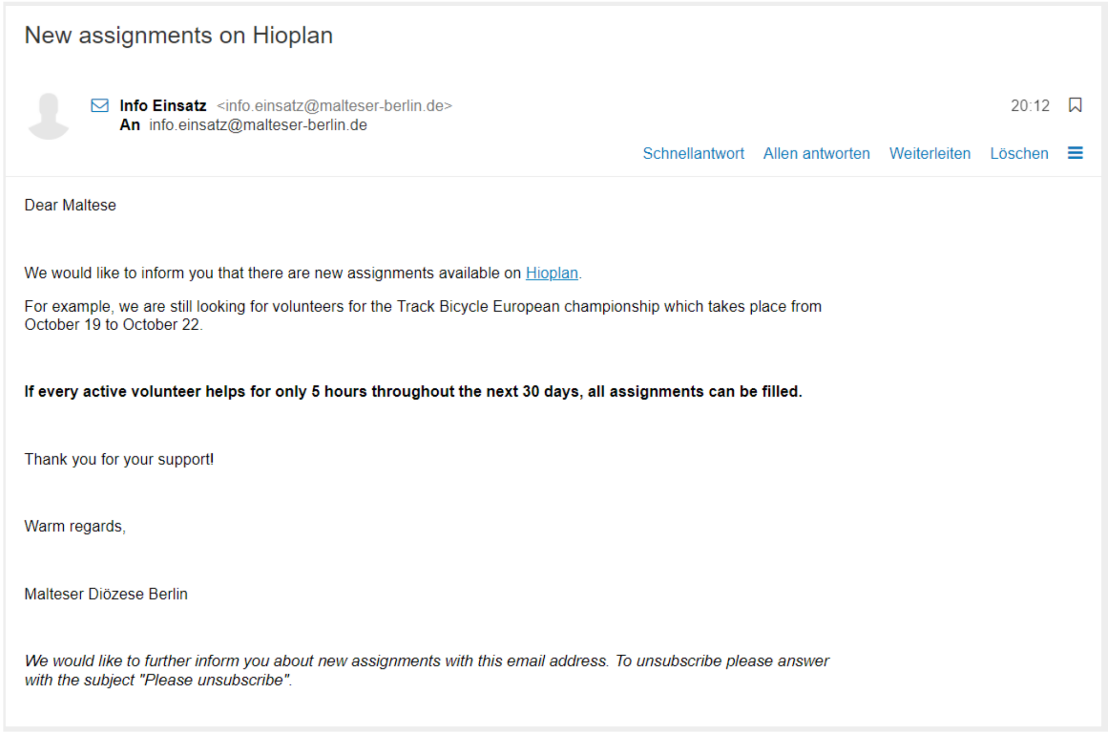
Both experimental groups received the newsletters three times, on October 9, October 23 and November 6, 2017. The average workload communicated in the three emails to the treatment group was 5.5 service hours (see Figures 3.6, 3.7, and 3.8 in Appendix A).

3.3.3 Field data

The Hioplan platform provides access to various event- and personel-related information. Our dataset contains all assignments between October 8, 2016 and November 19, 2017 that were performed by the volunteers of our sample.[46] In the following analyses, we consider the time frame from October 9, 2017 until November 19, 2017 as intervention

[45] Physicians were removed from the sample because their work is often paid and thus not voluntary. We also did not consider former volunteers with an inactive status on Hioplan.

[46] For reasons of time and administration, we were not able to track the assignments until December 3, 2017 (i.e., 30 days after the last newsletter) as originally planned.


New assignments on Hioplan

Info Einsatz <info.einsatz@malteser-berlin.de> 20:12
An info.einsatz@malteser-berlin.de

Schnellantwort Allen antworten Weiterleiten Löschen

Dear Maltese

We would like to inform you that there are new assignments available on Hioplan.

For example, we are still looking for volunteers for the Track Bicycle European championship which takes place from October 19 to October 22.

If every active volunteer helps for only 5 hours throughout the next 30 days, all assignments can be filled.

Thank you for your support!

Warm regards,

Malteser Diözese Berlin

We would like to further inform you about new assignments with this email address. To unsubscribe please answer with the subject "Please unsubscribe".

FIGURE 3.1: Translated newsletter example

period, while the year before the experiment serves as our baseline. As indicated in Figure 3.2, we refer to the six weeks before the experiment as the pre-intervention period. This pre-intervention period is used as a comparative time frame in our difference-in-difference analyses, whereas the longer baseline period is used to measure the volunteers' general level of engagement (see Sections 3.4.1 and 3.4.2).

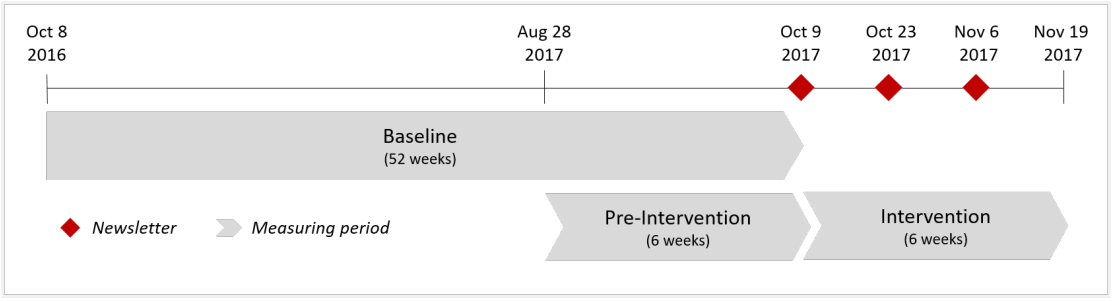


FIGURE 3.2: Measuring periods

In line with the selection of participants, we consider only care and first-aid assignments. Emergency services, which require specific paramedic skills, as well as training assignments that have a different character than the other voluntary services, were excluded. We further dropped assignments performed by three volunteers who unsubscribed from the newsletters.[47] Six volunteers and their corresponding assignments were disqualified

[47] These were two volunteers of the control and one of the treatment group.

because their email address was invalid, and one member was omitted because he was responsible for the data transfer and involved in the experimental procedure. This leads to a final sample of 2,966 assignment observations performed by 224 volunteers who actively took part in one or multiple voluntary services between October 8, 2016 and November 19, 2017. Of the 2,966 assignments, 2,682 took place in the baseline period, 333 in the pre-intervention period and 284 during the intervention.

As all assignments must be staffed, we are not able to investigate event-related characteristics that might influence volunteers' willingness to work. However, we can evaluate our treatment effect by comparing the engagement of the treatment and control groups within a certain time frame (i.e., the intervention period). In the following analyses, we focus on the volunteer-level of observation and use the hours of service per volunteer as our main dependent variable. The number of service hours is directly related to the performance information provided to the treatment group and is a more detailed outcome measure than the number of assignments, as certain assignments take place over multiple days. In our sample, one assignment takes on average 7 hours of service. A volunteer would therefore fulfil the required workload indicated in the newsletters with approximately one assignment per month.

Table 3.1 provides the observed volunteer characteristics of the final sample before the intervention. Besides the number of performed assignments and service hours, we have information on the volunteers' ages, divisions, and the date when their Hioplan account was created (i.e., tenure). Differences in continuous variables between the experimental groups were tested for significance using a two-sided t-test of equality of means. Differences in categorical variables were tested for significance using a $\tilde{\chi}^2$-test. Although our final sample includes only 224 of the initial participants, the p-values in the last column indicate that the average volunteer does not differ in terms of the observed characteristics across treatments. In particular, there are no significant differences between the treatment and control groups in the baseline or pre-intervention period with regard to the number of assignments or service hours per volunteer.

3.4 Results

3.4.1 Impact on volunteers' commitment

We first test our hypothesis that the performance information motivates volunteers to render more service hours. Note that we can compare the number of working hours between the treatment and control groups, although the overall workload is externally

TABLE 3.1: Sample characteristics and randomization check

	Sample n= 224	Control n=112	Treatment n=112	p-value
Berlin-North	0.22	0.23	0.21	0.629
Berlin-South	0.39	0.38	0.39	0.891
Berlin-West	0.39	0.38	0.4	0.785
Age (years)	28.52	28.62	28.42	0.904
	(12.15)	(12.04)	(12.32)	
Tenure (years)	2.49	2.46	2.52	0.949
	(2.06)	(2.05)	(2.08)	
Baseline assignments	11.97	11.54	12.40	0.630
	(17.54)	(15.68)	(19.29)	
Baseline service hours	84.87	82.90	86.85	0.73
	(121.15)	(109.04)	(132.62)	
Pre-intervention assignments	1.49	1.62	1.36	0.385
	(2.22)	(2.32)	(2.13)	
Pre-intervention service hours	12.37	13.08	11.66	0.389
	(17.49)	(18.05)	(16.91)	

Notes: The table reports the descriptive statistics of the final sample and for each treatment group individually. For categorical variables, the p-value in the last column is obtained from a $\tilde{\chi}^2$-test across the treatment and control groups. For continuous variables, the p-value is obtained from a two-sided t-test.

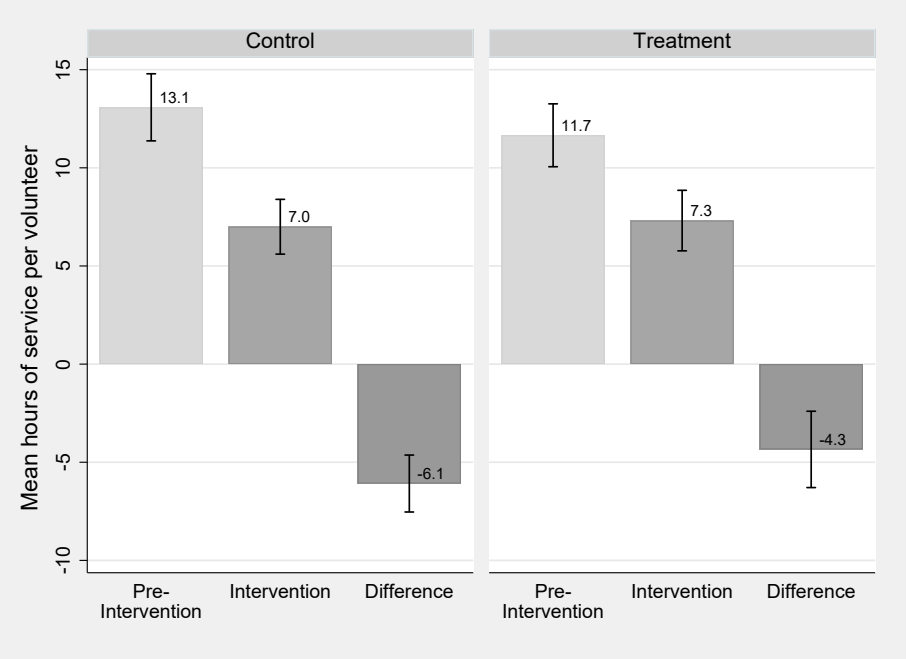
fixed. Figure 3.3 shows a descriptive analysis of the mean hours of service per volunteer for both experimental conditions in the pre-intervention and intervention periods.[48]

As the graph shows, there was a general decline in the number of working hours from the pre-intervention to the intervention period, independent of the experimental condition. The mean hours of service are only slightly higher in the treatment group than in the control group during the intervention period. However, these differences increase when taking into account the pre-intervention levels of performance. The change from the pre-intervention to the intervention period is approximately 1.8 hours higher in the treatment than in the control group.

We test the statistical significance of these differences in two regression models. We first focus on the intervention period and compare the hours of service in the treatment group to the control group in a simple OLS regression model of the following form:

$$\begin{aligned} y_i \ = \beta_0 + \beta_1 Treatment_i + \beta_2 Division_i + \beta_3 Baseline_i \\ + \beta_4 Age_i + \beta_5 Age_i^2 + \epsilon_i, \end{aligned} \tag{3.1}$$

[48] Recall that our outcome measure depends on the length of the observation period. We therefore suggest that the six-week pre-intervention period allows a more meaningful comparison of the activity levels before and during the intervention than the whole baseline.



Notes: Each bar indicates the mean hours of service performed per volunteer. The error bars represent the mean ± the standard error of the mean.

FIGURE 3.3: Hours of service before and during the intervention

where y_i is the number of service hours performed by volunteer i during the intervention period. Our main coefficient of interest, $Treatment_i$, indicates whether volunteer i was in the treatment (1) or control group (0). $Division_i$ is a dummy variable for a volunteer's divisional affiliation, that is, Berin-North, -South or -West. $Baseline_i$ refers to the number of service hours performed by volunteer i during the baseline period. As we expect that voluntary activity is also associated with a person's stage of life, we also control for Age_i and its quadratic function.[49] The random error term (ϵ_i) captures any unmodeled effects.

In a second model, we additionally control for volunteer-related differences between the treatment and control groups in the pre-intervention period. Here, we estimate the hours of service per volunteer in the following difference-in-difference analysis:

$$y_{i,t} = \beta_0 + \beta_1 Treatment_i + \beta_2 Period_t + \beta_3 (Treatment_i * Period_t) + \epsilon_{i,t}, \qquad (3.2)$$

where $y_{i,t}$ is the number of hours performed by volunteer i in period t. The variable $Period_t$ captures the time trend, indicating the pre-intervention (0) or intervention period

[49] We do not consider a volunteer's tenure at Malteser Berlin, because this measure is incomplete for 35 participants in our final sample. Additional analyses show that none of the results significantly change when $Tenure$ is included. Unfortunately, we have no information on the volunteers' gender to investigate potential gender effects.

(1). Our main coefficient of interest is the interaction term of the period dummy and the treatment group. It presents the expected mean difference in the hours of service, before and after the intervention between the treatment and control groups. Table 3.2 reports the estimates of Models 3.1 and 3.2, including robust and cluster-robust standard errors.

TABLE 3.2: Estimates of the commitment effect: Hours of service per volunteer

	Model 3.1		Model 3.2	
	1	2	3	4
Treatment	0.315	-0.074	-1.424	-1.849
	(2.080)	(1.708)	(2.345)	(1.797)
Berlin-South		1.114		0.395
		(1.292)		(1.279)
Berlin-West		4.467**		4.264**
		(1.976)		(1.804)
Baseline hours		0.076***		0.086***
		(0.009)		(0.007)
Age		0.002		-0.042
		(0.365)		(0.300)
Age^2		-0.000		0.000
		(0.005)		(0.004)
Period			-6.085***	-6.085***
			(1.454)	(1.462)
Treatment x Period			1.739	1.739
			(1.945)	(1.957)
Observations	224	224	448	448
N volunteers	224	224	224	224
R^2	0.000	0.354	0.025	0.407

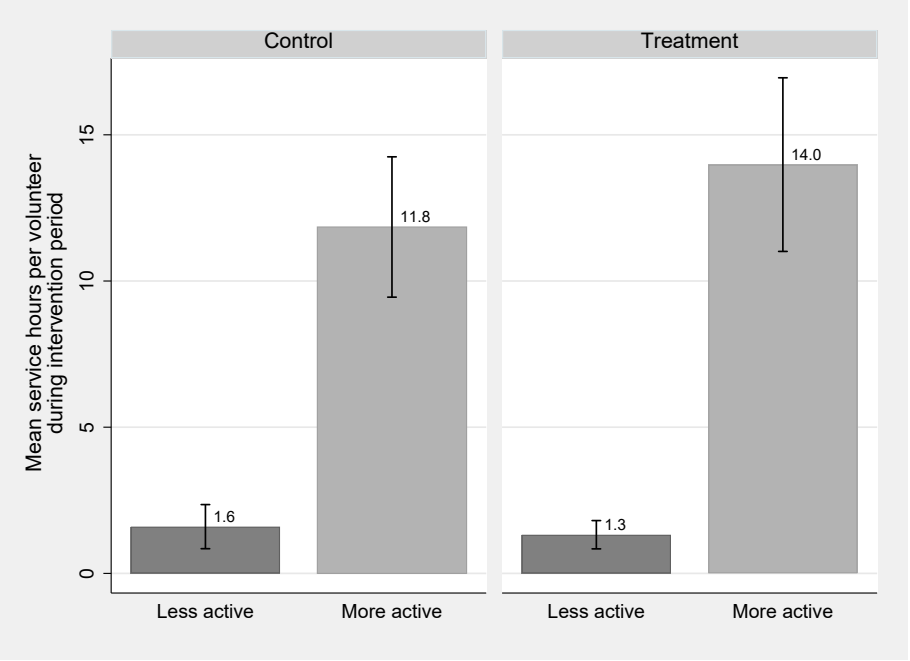
Notes: Specifications 1 and 2 present the results of an OLS regression with robust standard errors in parentheses. Specifications 3 and 4 are difference-in-difference regressions with cluster-robust standard errors in parentheses. The dependent variable is the number of service hours performed per volunteer. In addition to the treatment dummy, Specifications 2 and 4 further include a volunteer's number of service hours in the baseline period, his or her division, and years of age. *Period* indicates the pre-intervention (0) or intervention period (1). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

While we observe a slightly higher level of engagement in the treatment group (Specification 1), this difference vanishes to zero when the control variables are included (see Specification 2). As already indicated in Figure 3.3, the treatment effect is slightly stronger in the difference-in-difference analyses of Model 3.2. In Specification 3, the temporal change is 1.7 service hours higher in the treatment condition than in the control group. However, this effect is not significant. As expected, adding the (time-invariant) control variables in Specification 4 does no affect this result. We also find a similar outcome when we estimate the probability that a volunteer in the control or treatment group registers for a certain assignment in a logistic regression analysis (see Table 3.7, Appendix B). We therefore reject our hypothesis that the performance information motivates volunteers to become more actively engaged. Whether our

intervention has an impact on volunteers with lower prior activity levels and, therefore, supports a more balanced distribution of the workload at Malteser Berlin is investigated in the next section.

3.4.2 Impact on the workload distribution

To analyze the impact of the performance information on less-engaged versus more-engaged volunteers, we split our sample into two groups of equal sizes: those members who used to perform more than the median number of hours during the baseline period ("more-active") and those who performed less ("less-active").[50] Figure 3.4 shows the descriptive statistics of the workload distribution across the experimental conditions during the intervention period.



Notes: Each bar indicates the mean hours of service performed per volunteer during the intervention period. The error bars represent the mean ± the standard error of the mean.

FIGURE 3.4: Engagement of more- and less-active volunteers

Contrary to our expectation, we do not observe a more equal distribution among the more- and less-active volunteers in the treatment group. Less-active volunteers who received the performance information tended to contribute even less, whereas the more-engaged volunteers of the treatment group carried a larger share of the workload than those in

[50] The mean baseline engagement is 14.2 hours of service for the less-active volunteers and 155.5 hours for the more-active members. This equals 1.2 or 13 service hours per month, respectively. The use of three or four performance groups (terciles or quartiles) does not change our results.

the control group. We further analyse these graphical impressions in an OLS regression model of the following form:

$$y_i = \beta_0 + \beta_1 Treatment_i + \beta_2 More\text{-}active_i + \beta_3 (Treatment_i * More\text{-}active_i) + \beta X_i' + \epsilon_i, \tag{3.3}$$

where y_i is the number of service hours per volunteer during the intervention period. As before, $Treatment_i$ indicates the treatment (1) or control group (0). The variable $More\text{-}active_i$ is a dummy indicator for whether volunteer i belongs to the less or more-active 50%, based on his or her baseline engagement. Our main coefficient of interest is the interaction term between the $Treatment_i$ and the $More\text{-}active_i$ dummies, indicating the difference of the treatment effect between the two performance quantiles. X_i' is a vector with the volunteer-related control variables, including the volunteer's divisional affiliation and age. The term ϵ_i describes the random error.

Table 3.3 provides the estimates of Model 3.3. As suggested by the descriptive analysis, the performance information has a rather negative impact on the contribution of less-engaged volunteers when compared to the treatment effect on more-active members. More-active volunteers of the treatment group perform up to 2.4 additional hours of service than the more-active volunteers of the control group. However, this effect is statistically not significant in either specification. This outcome does not change if we use the number of service hours during the baseline period, instead of the 50% quantile, as a continuous mediator variable (see Table 3.8, Appendix B). We also obtain a similarly insignificant result when considering the workload distribution among the more- and less-active volunteers in the pre-intervention period. As shown in the difference-in-difference regression in Table 3.9 of Appendix B, the interaction term between the treatment group, the *More-active* dummy, and the intervention period is positive but not significant. Table 3.10 in Appendix B further confirms these results in a logistic regression model in which we consider each service as a single observation and estimate the chance that a more- or less-active volunteer performs a certain assignment. We therefore reject our hypothesis that the workload information motivates less-active volunteers to engage in more volunteer services.

TABLE 3.3: OLS regression: Treatment–performance interaction

	1	2
Treatment	-0.277	-0.790
	(0.896)	(0.976)
More-active	10.248***	9.923***
	(2.517)	(2.518)
Treatment x More-active	2.411	3.083
	(3.921)	(3.952)
Berlin-South		5.962***
		(1.871)
Berlin-West		6.308***
		(2.283)
Age		-0.034
		(0.387)
Age^2		-0.001
		(0.005)
Observations	224	224
R^2	0.138	0.174

Notes: The table presents the results of an OLS regression with robust standard errors in parentheses. The dependent variable is the number of service hours performed per volunteer. The variable *Treatment* indicates whether a volunteer belongs to the treatment (1) or control group (0). *More-active* is a dummy variable equal to 1 for the more-active 50% of the members and 0 otherwise. Specification 2 further includes the volunteer's division and years of age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

For a more detailed analysis of the distribution effect, we compare two measures of variation in the treatment and control groups. For each group, Table 3.4 reports the variance and the coefficient of variance for the hours of service during the intervention period.[51]

TABLE 3.4: Measures of variation per group

	Treatment	Control	Sample
Variance	266.42	218.12	241.21
Coefficient of Variance	2.23	2.11	2.17

In line with the regression results, both measures of inequality are higher in the treatment than in the control group. We can test the statistical significance of these differences with

[51] The coefficient of variance is defined as the ratio between the standard deviation and the mean. It is therefore scale-invariant and allows a comparison of the treatment and control groups in the case of different means (Schwartz and Winship 1980, Heshmati 2004).

a Brown-Forsythe test, which is a modified Levene's test for homogeneity of variances.[52] The Brown-Forsythe test does not support a differing workload distribution in the control and treatment group during the intervention period (p=0.905).[53] We therefore reject the hypothesis that our intervention leads to a more balanced workload distribution among the volunteers at Malteser Berlin. As indicated by the regression results, the performance information appears to encourage already-active members, rather than less-active volunteers. However, this effect does not induce a significant increase in inequality.

While we do not find a motivational or distributional effect of our intervention, we finally aim to investigate whether the newsletters themselves could have a positive impact on balancing the voluntary engagements. The next section provides an analysis of the potential reminder effect of the newsletters on the workload distribution at Malteser Berlin.

3.4.3 Potential reminder effects

By informing volunteers about the availability of new assignments, the newsletters of the treatment and control groups also constitute a simple reminder. Reminders have been shown to trigger behavioural changes in various settings (e.g., Apesteguia et al. 2013, Altmann and Traxler 2014, Bruhin et al. 2015), and their effect seems so be especially pronounced for individuals who show below-average levels of the desired behavior (Calzolari and Nardotto 2016). Therefore, the newsletters in our intervention may have encouraged less-active volunteers to sign up for more service hours across both experimental groups. In addition to the hypotheses tests in Sections 3.4.1 and 3.4.2, we investigate this potential reminder effect by a simple pretest–posttest analysis.[54] Specifically, we compare the engagement of the more- and less-active volunteers between the pre- and intervention period, independent of their assignment to the treatment or control group. As the total number of service hours is externally fixed and differs between the pre-intervention and intervention periods, we use the share on the total hours of

[52] Like the Levene's test, the Brown-Forsythe test examines the null hypothesis that there is homoscedasticity among two groups in the overall population. However, it uses the median, instead of the sample mean, to evaluate the dispersion. The Brown-Forsythe test is therefore more robust than the standard Levene's test to outliers and provides a better fit when values are not normally distributed, as in our case (Brown and Forsythe 1974).

[53] Box plots indicate that the dispersion in the treatment group is predominantly driven by one highly engaged volunteer, who performed more than 20 services during the intervention period. Excluding this outlier decreases the variance and the coefficient of variance in the treatment group but does not significantly change the variance test or the regression results.

[54] This extension lacks a clean control group, and we are aware of the threats to the internal validity of such a design (see Reichardt 2009, pp. 47–48).

service per volunteer as our main outcome measure. We estimate this variable in the following difference-in-difference model:

$$y_{i,t} = \beta_0 + \beta_1 Period_t + \beta_2 More\text{-}active_i + \beta_3 (Period_t * More\text{-}active_i) + \beta X_i' + \epsilon_{i,t}, \tag{3.4}$$

where $y_{i,t}$ is the number of service hours performed by volunteer i as a percentage of the total number of working hours in period t. As before, $Period_t$ shows the pre-intervention (0) or intervention period (1), and $More\text{-}active_i$ is a dummy variable indicating whether a volunteer belongs to the less- (0) or more-active 50% of the members (1). The term X_i' is a vector with additional, volunteer-related control variables, and $\epsilon_{i,t}$ indicates the random error. Table 3.5 provides the estimates of Model 3.4, using cluster-robust standard errors.

TABLE 3.5: Pretest–posttest regression: Share of the total service hours

	1	2
Period	-0.021	-0.021
	(0.033)	(0.033)
More-active	0.669***	0.674***
	(0.072)	(0.073)
Period x More-active	0.043	0.043
	(0.102)	(0.103)
Berlin-South		0.289***
		(0.087)
Berlin-West		0.317***
		(0.104)
Age		-0.005
		(0.015)
Age^2		-0.000
		(0.000)
Observations	448	448
N volunteers	224	224
R^2	0.179	0.209

Notes: The table presents the results of a difference-in-difference regression for the more- and less-active volunteers before and after the intervention. Standard errors clustered on the individual level are in parentheses. The dependent variable is the share of the total service hours per period performed per volunteer. The variable *Period* indicates the pre-intervention (0) or intervention period (1). *More-active* is a dummy variable equal to 1 for the more-active 50% of the volunteers, based on their baseline engagement. Specification 2 further includes the volunteer's division and age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Contrary to our expectation, the more-active volunteers increase their shares of the total service hours after the introduction of the newsletters compared to the less-active

volunteers.[55] This interaction is, however, not significant. As shown in Table 3.11 of Appendix B, the interaction term becomes significant at the 5% level if we use the number of service hours in the baseline period, instead of the *More-active* dummy, as a continuous mediator variable.

To further clarify these results, Table 3.6 provides the variance and the coefficient of variance for the hours of service per volunteer before and during the intervention. As already suggested by the regression results, both measures are higher after the introduction of the newsletters during the intervention period than in the pre-intervention period. The Brown-Forsythe test reports a significant difference in the variation between the two periods (p=0.000).

TABLE 3.6: Measures of variation per period

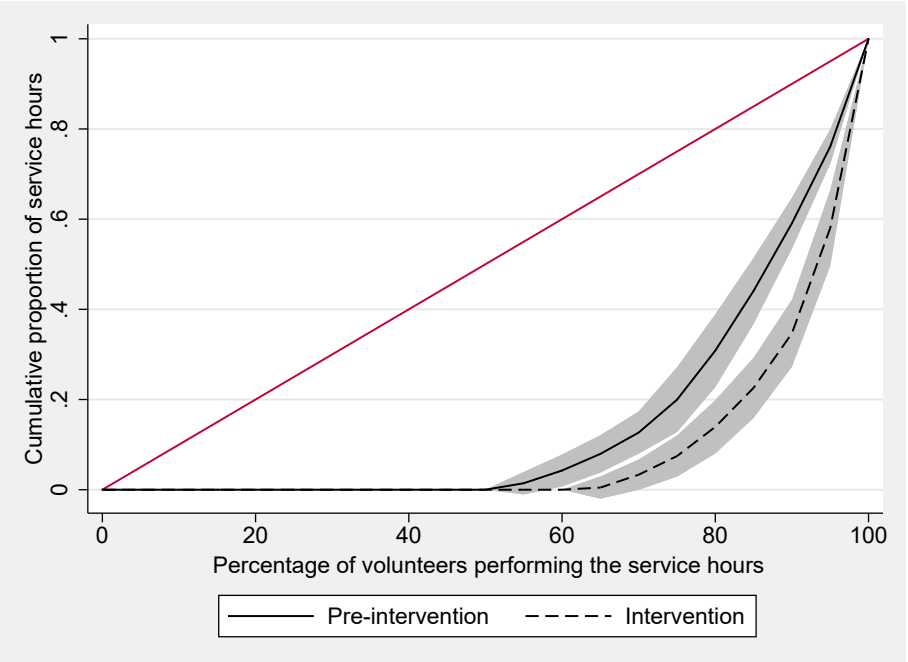
	Pre-Intervention	Intervention	Total
Variance	241.21	306.43	280.02
Coefficient of Variance	1.42	2.17	1.71

This finding is also supported by the Lorenz curves in Figure 3.5. The deviation of the Lorenz curves from the linear line of perfect equality is significantly larger for the intervention period, indicating that the inequality of the workload distribution increased during that time. The dotted line shows that approximately 60% of the volunteers did not participate in any assignment during the intervention period, while the total number of service hours was carried by the other 40% of the members.

3.5 Discussion

In this study, we investigated whether simple information on the average engagement needed per volunteer could motivate less-active members of a charitable organization to increase their voluntary commitment. In contrast to prior works that confirmed the impact of normative information on charitable contributions (Frey and Meier 2004, Shang and Croson 2009), we find no evidence for such a conformity effect. Our results show a slight increase in the number of voluntary service hours performed by the treatment group in comparison to the control group. This tendency is, however, not significant and mainly driven by volunteers with an above-average level of engagement. These volunteers seem to be more receptive to the internal performance information used in our intervention than less-active volunteers. The inequality of the service distribution at Malteser Berlin

[55] The interaction coefficient indicates that the more-active 50% raise their shares on average by 0.043 percentage points. In relative terms, this equals an increase of 6%.



Notes: The Lorenz curve illustrates the inequality of the workload distribution by showing the proportion of the overall hours of service (y-axis) assumed by the bottom "x" percent of the volunteers (x-axis). The gray-shaded areas reflect the 95% confidence intervals with robust standard errors clustered on the individual level.

FIGURE 3.5: Lorenz curves for the hours of service

therefore tends to increase. Although the effect is statistically not significant, this result contradicts existing literature around norm conformity (e.g., Cialdini and Goldstein 2004), suggesting that volunteers would adapt their effort toward the average required contribution per volunteer.

We explain these outcomes with several particularities of our field setting. First, the normative information used in our intervention differs from earlier studies (e.g., Chen et al. 2010) in the sense that is does not refer to the effective contribution levels of a peer group but is a hypothetical, future-related average. This possibly reduces the social sense of responsibility transmitted through the information. In line with earlier studies reporting a significant performance impact of relative feedback (e.g., Hannan et al. 2008, Azmat and Iriberri 2010, Blanes i Vidal and Nossol 2011), a more specific comparison of a volunteer's personal engagement and the average contributions of other members might have increased the effectiveness of our intervention. Furthermore, this project focused on existing members of a voluntary aid organization. The social ties between these participants bring the risk of contamination, which we acknowledge as a major limitation of our design. Last, following the results of Chen et al. (2010), we suggest that performance information can trigger competitive preferences and therefore reinforce the motivation of volunteers with an above-average engagement. For members who feel strongly committed to Malteser Berlin and its purpose, our intervention possibly promoted the number of working hours as a personal performance indicator and, thus,

encouraged self-monitoring and self-set performance goals. Both constitute a crucial source of self-motivation (Bandura 1991, Goerg and Kube 2012).

Additional analyses confirm that less-active volunteers are not only less responsive to the performance information but also show smaller reactions to the newsletters in general. Across both experimental groups, we find that members with a lower (higher) baseline engagement decrease (increase) their shares of the total service hours during the intervention period. The workload distribution therefore becomes significantly less equal compared to the pre-intervention period. To explain this result, we assume that the newsletters are interpreted as a sign of staff shortage. In this case, it does not seem surprising that highly committed volunteers show a stronger response. In a similar context, Bruhin et al. (2015) demonstrate that phone calls informing voluntary blood donors of a current shortage of their blood type significantly increases their contributions. The short-term effect is especially strong for highly motivated donors, who exhibit a high baseline donation rate. Our result support this idea that organizational commitment crucially influences the effectiveness of internal communication strategies in charitable organizations. However, due to the methodological limitations of the pretest–posttest design, this finding requires further validation.[56]

How to motivate less-engaged volunteers to reach a more balanced distribution of volunteer assignments remains an open question for future research. Aside from testing different types of normative information, it would be interesting to explore the role of altruistic and competitive preferences driving volunteers' reactions to such information. Future studies may also look at other non-monetary incentives to increase volunteers' motivation. Given the strong motivating effect of social recognition demonstrated by Kosfeld and Neckermann (2011), forthcoming studies may want to investigate symbolic awards in the context of volunteering. The fact that effective motivational practices from paid employment are not directly transferable to the voluntary sector leaves various questions to future research.

[56] Because we lack of a clean comparison group, we can, for example, not control for the possibility that more-active and less-active volunteers respond differently to workload peaks. More-active volunteers may keep a high level of engagement over time, whereas less-active volunteers may only increase their contribution when staff are critically short. The higher workload of the pre-intervention period could therefore support a more equal distribution of the service hours in comparison to the less-busy intervention period.

References

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., and Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology*, 29(2):97–119.

Allen, N. J. and Rushton, J. P. (2016). Personality characteristics of community mental health volunteers: A review. *Journal of Voluntary Action Research*, 12(1):36–49.

Allen, V. L. (1965). Situational factors in conformity. In Berkowitz, L., editor, *Advances in experimental social psychology*, volume 2, pages 133–175. Academic Press, New York.

Altmann, S. and Traxler, C. (2014). Nudges at the dentist. *European Economic Review*, 72:19–38.

Anderson, J. C. and Moore, L. F. (2016). The motivation to volunteer. *Journal of Voluntary Action Research*, 7(3-4):120–129.

Apesteguia, J., Funk, P., and Iriberri, N. (2013). Promoting rule compliance in daily-life: Evidence from a randomized field experiment in the public libraries of barcelona. *European Economic Review*, 64:266–284.

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

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2):191–215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2):122–147.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2):248–287.

Barnes, M. L. and Sharpe, E. K. (2009). Looking beyond traditional volunteer management: A case study of an alternative approach to volunteer engagement in parks and recreation. *International Journal of Voluntary and Nonprofit Organizations*, 20(2):169–187.

Bennett, R. and Barkensjo, A. (2005). Internal marketing, negative experiences, and volunteers'commitment to providing high-quality services in a uk helping and caring charitable organization. *International Journal of Voluntary and Nonprofit Organizations*, 16(3):251–274.

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

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Boezeman, E. J. and Ellemers, N. (2008). Volunteer recruitment: The role of organizational support and anticipated respect in non-volunteers' attraction to charitable volunteer organizations. *Journal of Applied Psychology*, 93(5):1013–1026.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367.

Bruhin, A., Goette, L., Roethlisberger, A., Markovic, A., Buchli, R., and Frey, B. M. (2015). Call of duty: The effects of phone calls on blood donor motivation. *Transfusion*, 55(11):2645–2652.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Bundesamt für Statistik [BFS] (2017a). *Freiwilligenarbeit.* Retrieved from https://www.bfs.admin.ch/bfs/de/home/statistiken/querschnittsthemen/wohlfahrtsmessung/alle-indikatoren/gesellschaft/freiwilligenarbeit.html.

Bundesamt für Statistik [BFS] (2017b). *Freiwilligenarbeit, Zeiteinsatz: Mittlere Stundenzahl pro Person und Monat.* Retrieved from: https://www.bfs.admin.ch/bfs/de/home/statistiken/arbeit-erwerb/unbezahlte-arbeit/freiwilligenarbeit.assetdetail.2922632.html.

Bussell, H. and Forbes, D. (2002). Understanding the volunteer market: The what, where, who and why of volunteering. *International Journal of Nonprofit and Voluntary Sector Marketing*, 7(3):244–257.

Buunk, B. P. and Mussweiler, T. (2001). New directions in social comparison research. *European Journal of Social Psychology*, 31(5):467–475.

Calzolari, G. and Nardotto, M. (2016). Effective reminders. *Management Science*, 63(9):2915–2932.

Carpenter, J. and Myers, C. K. (2010). Why volunteer? evidence on the role of altruism, image, and incentives. *Journal of Public Economics*, 94(11-12):911–920.

Chen, Y., Harper, F. M., 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.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015–1026.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In Gilbert, T., Fiske, S. T., and Lindzey, G., editors, *The handbook of social psychology*, pages 151–192. McGraw-Hill, New York, NY, US.

Clary, E. G. and Miller, J. (1986). Socialization and situational influences on sustained altruism. *Child Development*, 57(6):1358–1369.

Clary, E. G., Snyder, M., Ridge, R. D., Copeland, J., Stukas, A. A., Haugen, J., and Miene, P. (1998). Understanding and assessing the motivations of volunteers: A functional approach. *Journal of Personality and Social Psychology*, 74(6):1516–1530.

Clary, E. G., Snyder, M., Ridge, R. D., Miene, P. K., and Haugen, J. A. (1994). Matching messages to motives in persuasion: A functional approach to promoting volunteerism1. *Journal of Applied Social Psychology*, 24(13):1129–1146.

Cnaan, R. A. and Cascio, T. A. (1998). Performance and commitment. *Journal of Social Service Research*, 24(3-4):1–37.

Cnaan, R. A. and Goldberg-Glen, R. S. (1991). Measuring motivation to volunteer in human services. *The Journal of Applied Behavioral Science*, 27(3):269–284.

Coleman, J. S. (2000). *Foundations of social theory*. Belknap Press of Harvard University Press, Cambridge, Mass.

Craig-Lees, M., Harris, J., and Lau, W. (2008). The role of dispositional, organizational and situational variables in volunteering. *Journal of Nonprofit & Public Sector Marketing*, 19(2):1–24.

Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326.

Deutsch, M. and Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3):629–636.

Eden, D. and Kinnar, J. (1991). Modeling galatea: Boosting self-efficacy to increase volunteering. *Journal of Applied Psychology*, 76(6):770–780.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fisher, R. J. and Ackerman, D. (1998). The effects of recognition and group need on volunteerism: A social norm perspective. *Journal of Consumer Research*, 25(3):262–275.

Frey, B. S. and Meier, S. (2004). Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, 94(5):1717–1722.

Frisch, M. B. and Gerrard, M. (1981). Natural helping systems: A survey of red cross volunteers. *American Journal of Community Psychology*, 9(5):567–579.

Gidron, B. (1978). Volunteer work and its rewards. *Volunteer Administration*, 11(3):18–32.

Goerg, S. J. and Kube, S. (2012). Goals (th)at work – goals, monetary incentives, and workers' performance. Working Paper Series of the Max Planck Institute for Research on Collective Goods 2012/19, Bonn.

Graf, S. (2015). Motivational and organizational factors for stimulating membership and donor engagement. Doctoral Thesis, University of Fribourg.

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

Heshmati, A. (2004). Inequalities and their measurement. IZA Discussion Paper No. 1219.

Hustinx, L. and Lammertyn, F. (2016). The cultural bases of volunteering: Understanding and predicting attitudinal differences between flemish red cross volunteers. *Nonprofit and Voluntary Sector Quarterly*, 33(4):548–584.

Jenner, J. R. (2016). Participation, leadership, and the role of vollinteerism among selected women volunteers. *Journal of Voluntary Action Research*, 11(4):27–38.

Keser, C. and van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102(1):23–39.

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

Latham, G. P. and Locke, E. A. (1991). Self-regulation through goal setting. *Organizational Behavior and Human Decision Processes*, 50(2):212–247.

Lindenmeier, J. (2008). Promoting volunteerism: Effects of self-efficacy, advertisement-induced emotional arousal, perceived costs of volunteering, and message framing. *International Journal of Voluntary and Nonprofit Organizations*, 19(1):43–65.

Locke, E. A., Latham, G. P., and Smith, K. J. (1990). *A theory of goal setting & task performance*. Prentice Hall, Englewood Cliffs, N.J.

Loewenstein, G. F., Thompson, L., and Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology*, 57(3):426–441.

Lott, A. J. and Lott, B. E. (1961). Group cohesiveness, communication level, and conformity. *The Journal of Abnormal and Social Psychology*, 62(2):408–412.

Machin, J. and Paine, A. E. (2008). *Management matters: A national survey of volunteer management capacity.* Institute for Volunteering Research, London.

Martinez, T. A. and McMullin, S. L. (2016). Factors affecting decisions to volunteer in nongovernmental organizations. *Environment and Behavior*, 36(1):112–126.

Penner, L. A. (2002). Dispositional and organizational influences on sustained volunteerism: An interactionist perspective. *Journal of Social Issues*, 58(3):447–467.

Penner, L. A. and Finkelstein, M. A. (1998). Dispositional and structural determinants of volunteerism. *Journal of Personality and Social Psychology*, 74(2):525–537.

Penner, L. A., Fritzsche, B. A., Craiger, J. P., and Freifeld, T. S. (1995). Measuring the prosocial personality. In Butcher, J. N. and Spielberger, C. D., editors, *Advances in personality assessment*, pages 147–163. Erlbaum, Hillsdale, NJ.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, 6(1):65–78.

Reichardt, C. S. (2009). Quasi-experimental design. In Maydeu-Olivares, A. and Millsap, R. E., editors, *The SAGE handbook of quantitative methods in psychology*, pages 46–71. SAGE, Los Angeles and London.

Roberts, R. D. (1984). A positive model of private charity and public transfers. *Journal of Political Economy*, 92(1):136–148.

Rosenhan, D. L. (1970). The natural socialisation of altruistic autonomy. In Macaulay, J. R. and Berkowitz, L., editors, *Altruism and helping behavior*, pages 251–268. Academic Press, New York.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Schwartz, J. and Winship, C. (1980). The welfare approach to measuring inequality. *Sociological Methodology*, 11:1.

Schwartz, S. H. (1965). Normative influences on altruism. In Berkowitz, L., editor, *Advances in experimental social psychology*, volume 10, pages 221–279. Academic Press, New York.

Shang, J. and Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119(540):1422–1439.

Simonson, J., Vogel, C., and Tesch-Römer, C. (2017). *Freiwilliges Engagement in Deutschland: Der Deutsche Freiwilligensurvey 2014*. Springer Fachmedien Wiesbaden, Wiesbaden.

Stajkovic, A. D. and Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, 124(2):240–261.

Stirling, C., Kilpatrick, S., and Orpin, P. (2011). A psychological contract perspective to the link between non-profit organizations' management practices and volunteer sustainability. *Human Resource Development International*, 14(3):321–336.

Studer, S. and von Schnurbein, G. (2013). Organizational factors affecting volunteers: A literature review on volunteer coordination. *International Journal of Voluntary and Nonprofit Organizations*, 24(2):403–440.

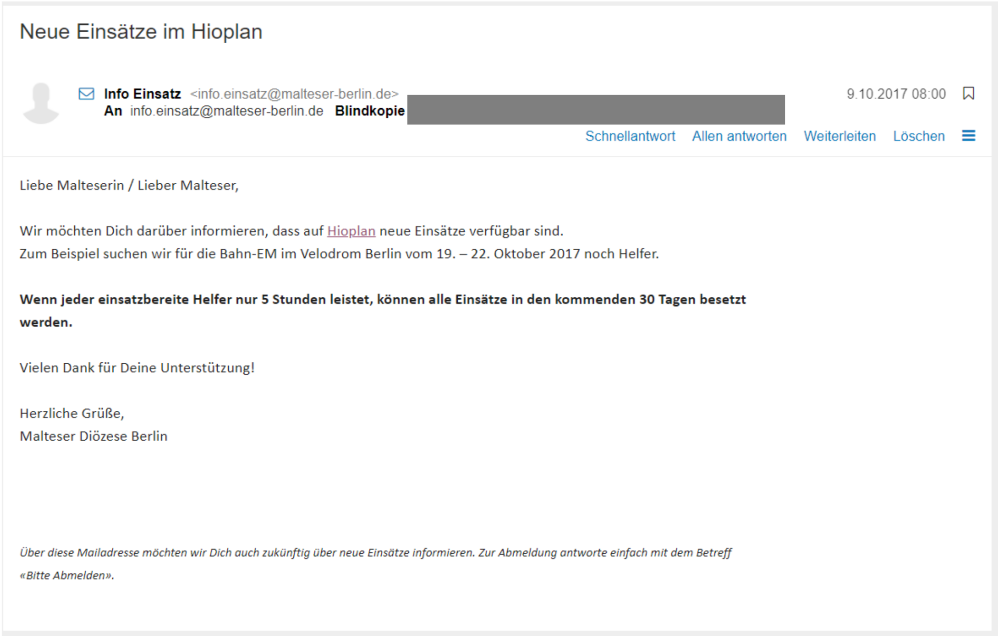
Suls, J., Martin, R., and Wheeler, L. (2016). Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science*, 11(5):159–163.

Vesterlund, L. (2003). The informational value of sequential fundraising. *Journal of Public Economics*, 87(3-4):627–657.

Warr, P. G. (1982). Pareto optimal redistribution and private charity. *Journal of Public Economics*, 19(1):131–138.

Ziemek, S. (2006). Economic analysis of volunteers' motivations—a cross-country study. *The Journal of Socio-Economics*, 35(3):532–555.

Appendix A Newsletters



Neue Einsätze im Hioplan

Info Einsatz <info.einsatz@malteser-berlin.de>
An info.einsatz@malteser-berlin.de **Blindkopie** [redacted]
9.10.2017 08:00

Schnellantwort Allen antworten Weiterleiten Löschen

Liebe Malteserin / Lieber Malteser,

Wir möchten Dich darüber informieren, dass auf Hioplan neue Einsätze verfügbar sind.
Zum Beispiel suchen wir für die Bahn-EM im Velodrom Berlin vom 19. – 22. Oktober 2017 noch Helfer.

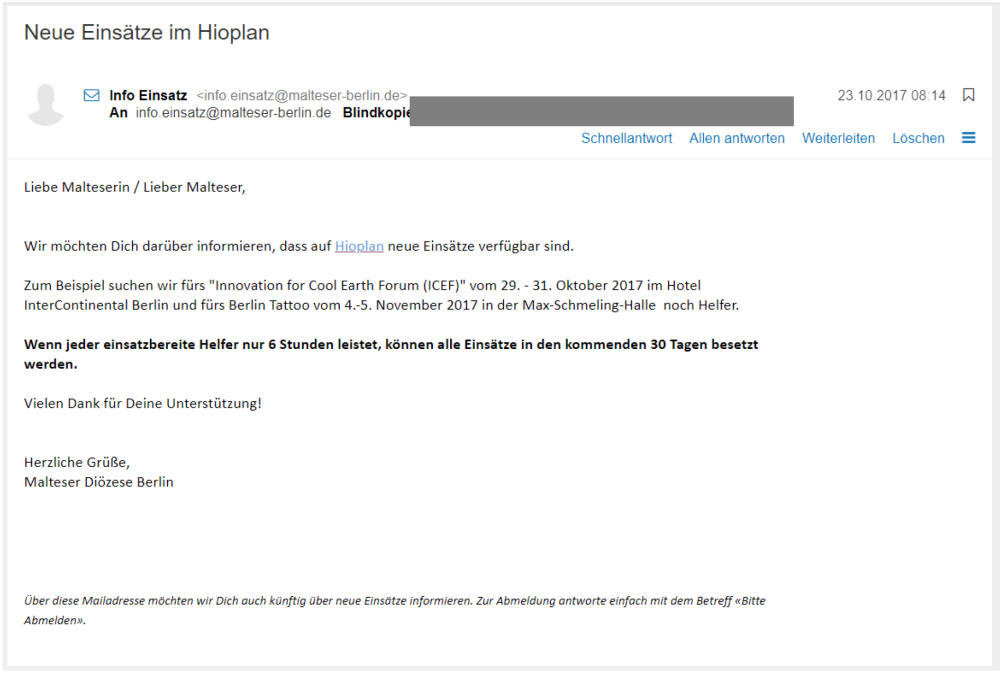
Wenn jeder einsatzbereite Helfer nur 5 Stunden leistet, können alle Einsätze in den kommenden 30 Tagen besetzt werden.

Vielen Dank für Deine Unterstützung!

Herzliche Grüße,
Malteser Diözese Berlin

Über diese Mailadresse möchten wir Dich auch zukünftig über neue Einsätze informieren. Zur Abmeldung antworte einfach mit dem Betreff «Bitte Abmelden».

FIGURE 3.6: Treatment newsletter 1



Neue Einsätze im Hioplan

Info Einsatz <info.einsatz@malteser-berlin.de>
An info.einsatz@malteser-berlin.de **Blindkopie** [redacted]
23.10.2017 08:14

Schnellantwort Allen antworten Weiterleiten Löschen

Liebe Malteserin / Lieber Malteser,

Wir möchten Dich darüber informieren, dass auf Hioplan neue Einsätze verfügbar sind.

Zum Beispiel suchen wir fürs "Innovation for Cool Earth Forum (ICEF)" vom 29. - 31. Oktober 2017 im Hotel InterContinental Berlin und fürs Berlin Tattoo vom 4.-5. November 2017 in der Max-Schmeling-Halle noch Helfer.

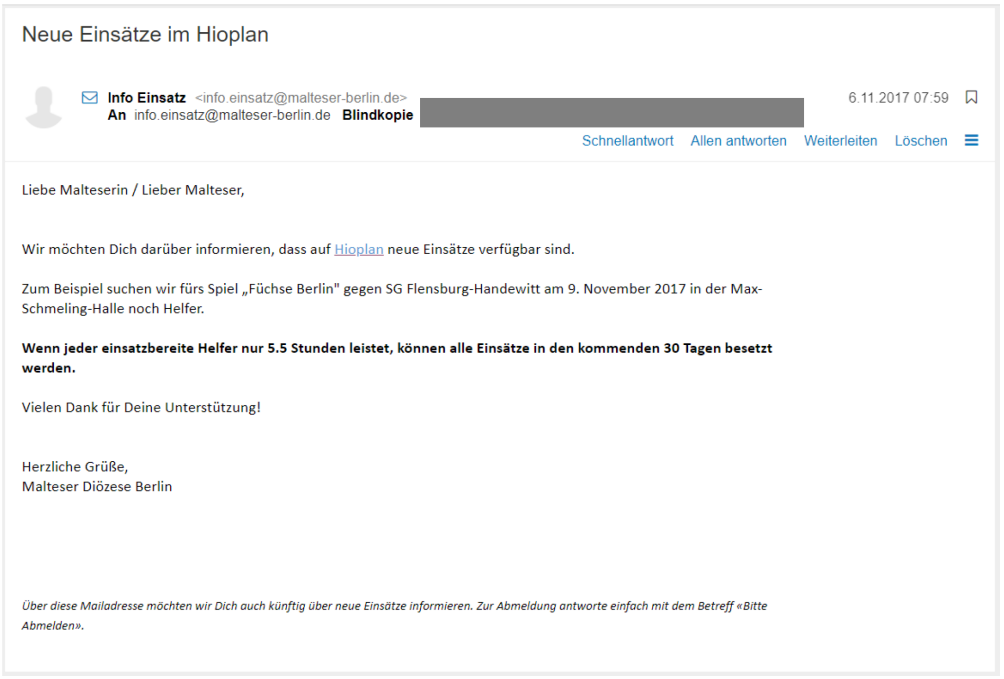
Wenn jeder einsatzbereite Helfer nur 6 Stunden leistet, können alle Einsätze in den kommenden 30 Tagen besetzt werden.

Vielen Dank für Deine Unterstützung!

Herzliche Grüße,
Malteser Diözese Berlin

Über diese Mailadresse möchten wir Dich auch künftig über neue Einsätze informieren. Zur Abmeldung antworte einfach mit dem Betreff «Bitte Abmelden».

FIGURE 3.7: Treatment newsletter 2


Neue Einsätze im Hioplan

Info Einsatz <info.einsatz@malteser-berlin.de>
An info.einsatz@malteser-berlin.de **Blindkopie** [redacted]

6.11.2017 07:59

Schnellantwort Allen antworten Weiterleiten Löschen

Liebe Malteserin / Lieber Malteser,

Wir möchten Dich darüber informieren, dass auf Hioplan neue Einsätze verfügbar sind.

Zum Beispiel suchen wir fürs Spiel „Füchse Berlin" gegen SG Flensburg-Handewitt am 9. November 2017 in der Max-Schmeling-Halle noch Helfer.

Wenn jeder einsatzbereite Helfer nur 5.5 Stunden leistet, können alle Einsätze in den kommenden 30 Tagen besetzt werden.

Vielen Dank für Deine Unterstützung!

Herzliche Grüße,
Malteser Diözese Berlin

Über diese Mailadresse möchten wir Dich auch künftig über neue Einsätze informieren. Zur Abmeldung antworte einfach mit dem Betreff «Bitte Abmelden».

FIGURE 3.8: Treatment newsletter 3

Appendix B Robustness checks

TABLE 3.7: Random effects logit regression: Odds that a volunteer participates

	1	2
Treatment	1.109	0.926
	(0.356)	(0.233)
Berlin-South		0.812
		(0.284)
Berlin-West		1.631
		(0.521)
Baseline hours		1.008***
		(0.001)
Age		0.942
		(0.058)
Age^2		1.000
		(0.001)
sd (volunteers)	1.770	1.217
Rho	0.488	0.310
Observations	63,616	63,616
N volunteers	224	224

Notes: The table presents the odds ratios of a logistic regression with random effects for volunteers. Robust standard errors clustered on the individual level are shown in parentheses. The dependent variable is a dummy variable indicating whether a volunteer engaged in an assignment. In addition to the *Treatment* dummy variable, Specification 2 includes the volunteer's number of service hours in the baseline period, his or her division and years of age. The total number of observations is 224 volunteers x 284 assignments in the intervention period = 63,616. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 3.8: OLS regression: Continuous performance–treatment interaction

	1	2
Treatment	-0.052	-0.076
	(0.265)	(0.265)
Baseline hours	0.012***	0.012***
	(0.003)	(0.003)
Treatment x Baseline hours	0.002	0.002
	(0.003)	(0.003)
Berlin-South		0.025
		(0.218)
Berlin-West		0.713**
		(0.353)
Age		-0.016
		(0.064)
Age^2		0.000
		(0.001)
Observations	224	224
R^2	0.344	0.362

Notes: The table presents the results of an OLS regression with robust standard errors provided in parentheses. The dependent variable is the number of service hours performed per volunteer during the intervention. The variable *Treatment* indicates whether a volunteer belongs to the treatment (1) or control group (0), and *Basline hours* shows a volunteer's number of service hours in the baseline period. Specification 2 further includes the volunteer's division and years of age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 3.9: Difference-in-difference regression of the distribution effect

	1	2
Treatment	-0.189	-0.269
	(0.164)	(0.187)
Intervention	-0.245	-0.245
	(0.150)	(0.151)
Treatment x Period	0.228	0.228
	(0.214)	(0.215)
More active	2.607***	2.564***
	(0.421)	(0.421)
Treatment x More-active	0.110	0.210
	(0.575)	(0.567)
Intervention x More-active	-0.856**	-0.856**
	(0.388)	(0.390)
Treatment x Period x More-active	0.307	0.307
	(0.577)	(0.580)
Berlin-South		0.967***
		(0.301)
Berlin-West		0.994***
		(0.341)
Age		-0.024
		(0.051)
Age^2		0.000
		(0.001)
Observations	448	448
N volunteers	224	224
R^2	0.208	0.240

Notes: The table presents the results of a difference-in-difference regression with robust standard errors clustered on the individual level in parentheses. The dependent variable is the number of service hours performed per volunteer. The variable $Treatment$ indicates whether a volunteer belongs to the treatment (1) or control group (0). *Period* is a dummy variable for the pre-intervention (0) or intervention period (1). The variable *Basline hours* shows a volunteer's number of service hours performed in the baseline period. Specification 2 further includes the volunteer's division and years of age. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 3.10: Random effects logit regression of the distribution effect

	1	2
Treatment	1.013	0.964
	(0.341)	(0.308)
Baseline hours	1.008***	1.008***
	(0.001)	(0.001)
Treatment x Baseline hours	1.000	1.000
	(0.001)	(0.001)
Berlin-South		0.811
		(0.284)
Berlin-West		1.631
		(0.521)
Age		0.942
		(0.058)
Age^2		1.000
		(0.001)
sd (volunteers)	1.281	1.257
Rho	0.333	0.324
Observations	63,616	63,616
N volunteers	224	224

Notes: The table presents the odds ratios of a logistic regression with random effects for volunteers. Robust standard errors clustered on the individual level are shown in parentheses. The dependent variable is a dummy variable indicating whether a volunteer engaged in an assignment. The variable *Treatment* indicates whether a volunteer belongs to the treatment (1) or control group (0), and *Basline hours* shows a volunteer's number of service hours in the baseline period. Specification 2 further includes the volunteer's division and years of age. The total number of observations is 224 volunteers x 284 assignments in the intervention period = 63,616. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 3.11: Pretest–posttest regression: Continuous performance interaction

	1	2
Period	-0.110**	-0.110**
	(0.048)	(0.048)
Baseline hours	0.003***	0.003***
	(0.000)	(0.000)
Period x Baseline hours	0.001**	0.001**
	(0.001)	(0.001)
Berlin-South		0.028
		(0.059)
Berlin-West		0.211**
		(0.088)
Age		-0.001
		(0.015)
Age^2		0.000
		(0.000)
Observations	448	448
N Volunteers	224	224
R^2	0.361	0.375

Notes: The table presents the results of an OLS regression with robust standard errors clustered on the individual level in parentheses. The dependent variable is the share of the total service hours per period performed per volunteer. The variable *Period* indicates the pre-intervention (0) or intervention period (1). *Basline hours* shows the number of service hours performed in the baseline period. Specification 2 further includes the volunteer's division and age. 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: 

Date: 17. Januar 2019