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

Angela Steffen, Frauke von Bieberstein †

Working Paper, February 21, 2019

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

*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.

†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.

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 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 3. Section 4 presents the empirical results. Finally, Section 5 discusses the findings and approaches for future research.

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 2.1 and 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 2.3.

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.

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 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 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 3.1). 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 3.2). Both characteristics should reinforce the positive impact of feedback on performance (see Kluger and DeNisi 1996).

[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).

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 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 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.*

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. Partici-

pants 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.*

3 Methodology

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

[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.

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 and Table 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).

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

[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.

[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

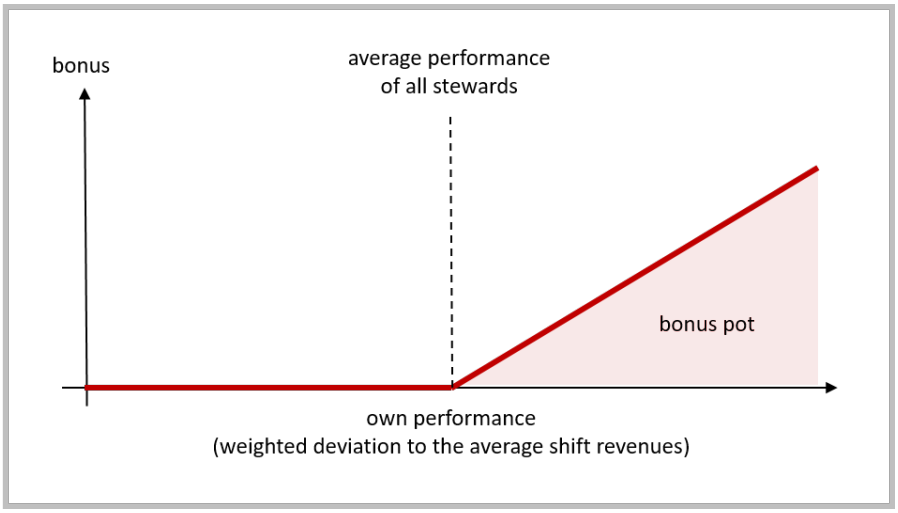


Figure 1: Incentive scheme

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.

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

deviation of 11.7 percentage points (see Figure 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.

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 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 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.

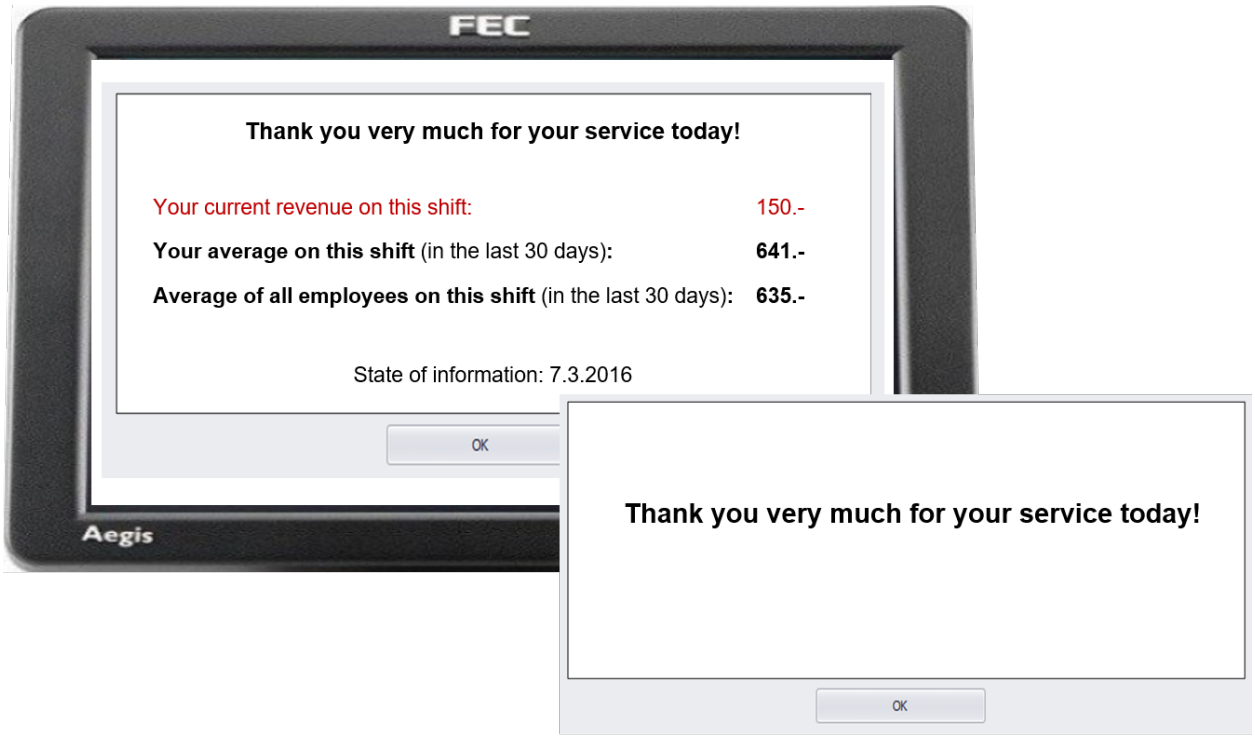


Figure 2: Translated message examples

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

[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.

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.

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

[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.

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

[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 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.

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 5 and 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.

4 Results

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

[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

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:

$$Ln(revhour)_{i,j,t} = \beta_0 + \beta_1 Group_i + \beta Stew'_i + \beta Shift'_j + \beta Date'_t + \beta Service'_{i,j,t} + \upsilon_i + \epsilon_{i,j,t}. \quad (1)$$

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

the service level.

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

service-related characteristics $Serv'$ that are shift- *and* date-specific. These variables include the average train occupancy of the service$_{i,j,t}$ 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 indicate random steward-specific deviations from the average (υ_i) and the random error ($\epsilon_{i,j,t}$).

Table 4 provides the estimates of Model 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.

[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 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.

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 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 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

Table 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 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 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 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 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 estimates in Table 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 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

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

[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.

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 3.1 and 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. As shown in Table 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.

4.2 Real-time feedback effects and ability

As set out in Section 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 4).[16] With reference to Model 1, we estimate the following interaction effects:

$$Log(revhour)_{i,j,t} = \beta_0 + \beta_1(Group_i * Quartile_i) + \beta Stew'_i + \beta Shift'_j + \beta Date'_t + \beta Service'_{i,j,t} + \upsilon_i + \epsilon_{i,j,t}. \quad (2)$$

$Log(revhour)_{i,j,t}$ 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 (see Specification 5 in Table 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 6 provides the estimates of Model 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 to the Wald tests, the

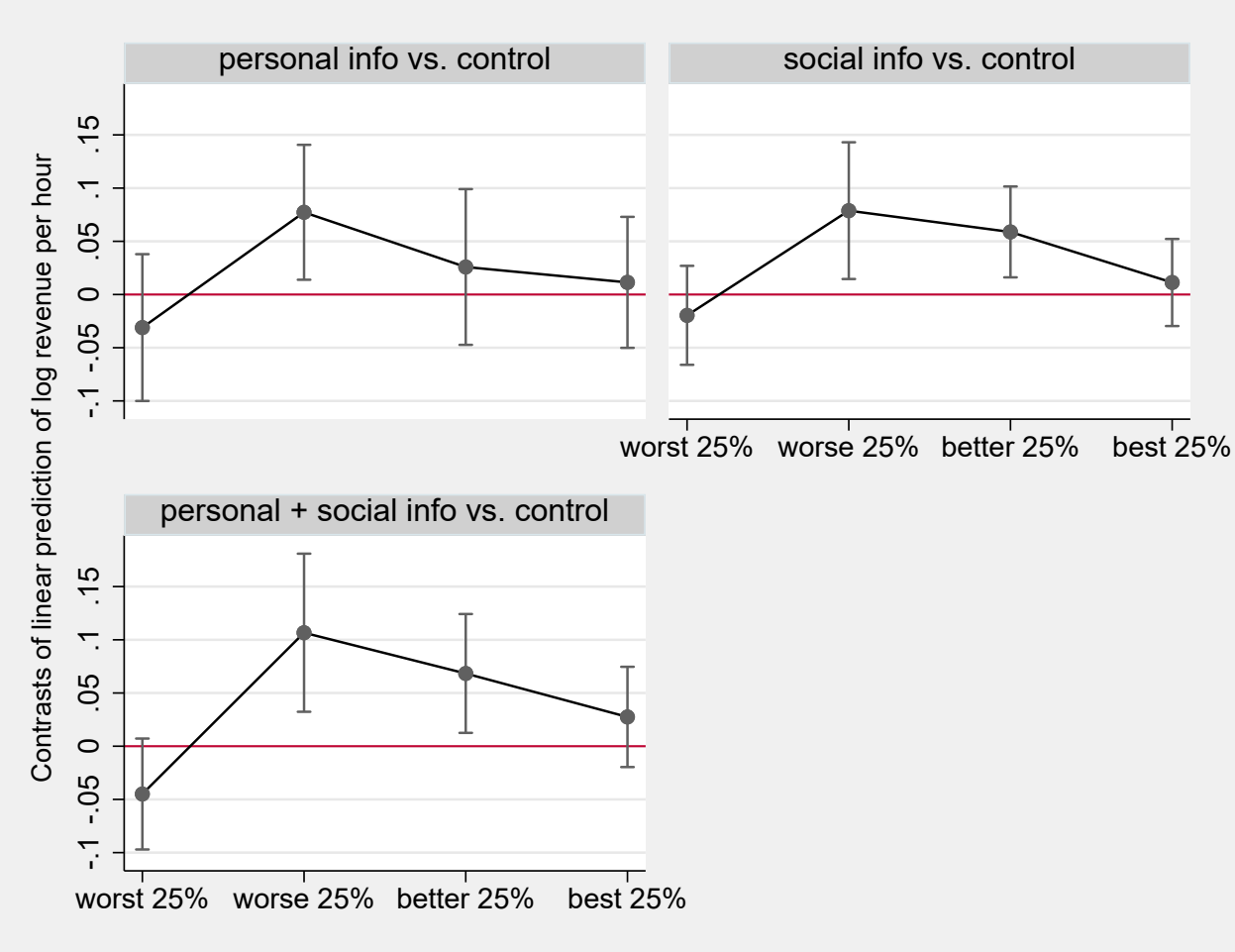
[16] Using this performance measure instead of a more recent or dynamic indicator allows us to uniquely assign each employee to one of the four performance groups and avoids endogenous interactions with our intervention.

Table 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 and 2 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 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 2) of the treatment groups compared to the control group. The error bars indicate the 95% confidence intervals.

Figure 3: Contrasts of predictive margins of Model 2

As indicated in Table 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 4.1 (see Figure 7 in Appendix B).

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

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

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.

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.

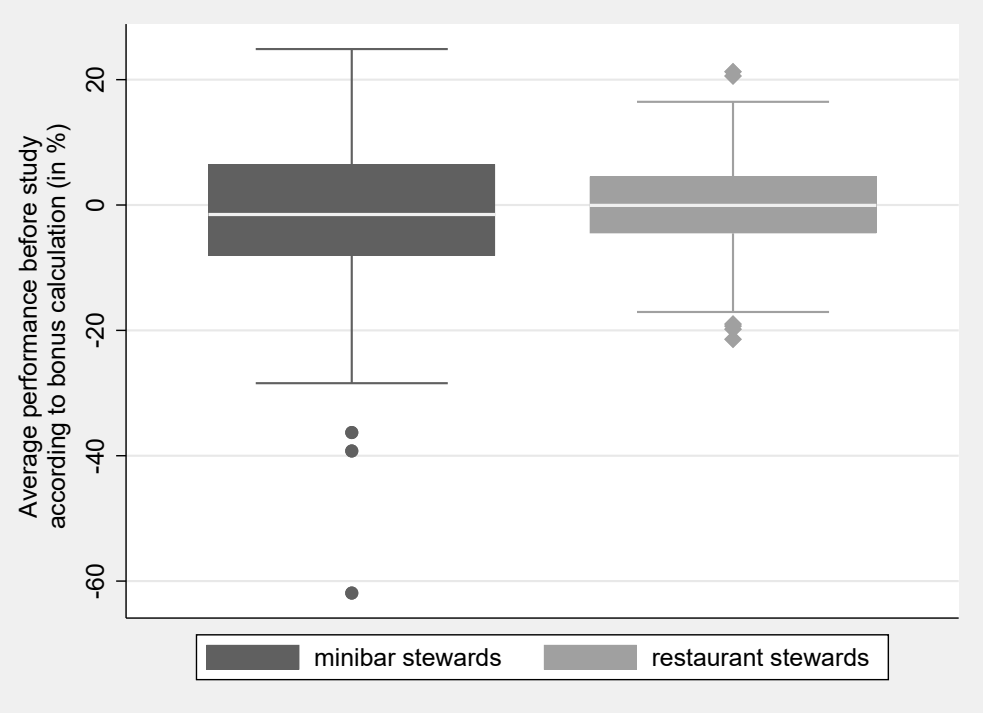
Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., and Staake, T. (2018). Overcoming salience bias: How real-time feedback fosters resource conservation. *Management Science*, 64(3):1458–1476.

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 3.1) over all months of the pre-study period.

Figure 4: Box plot of the pre-study steward performance

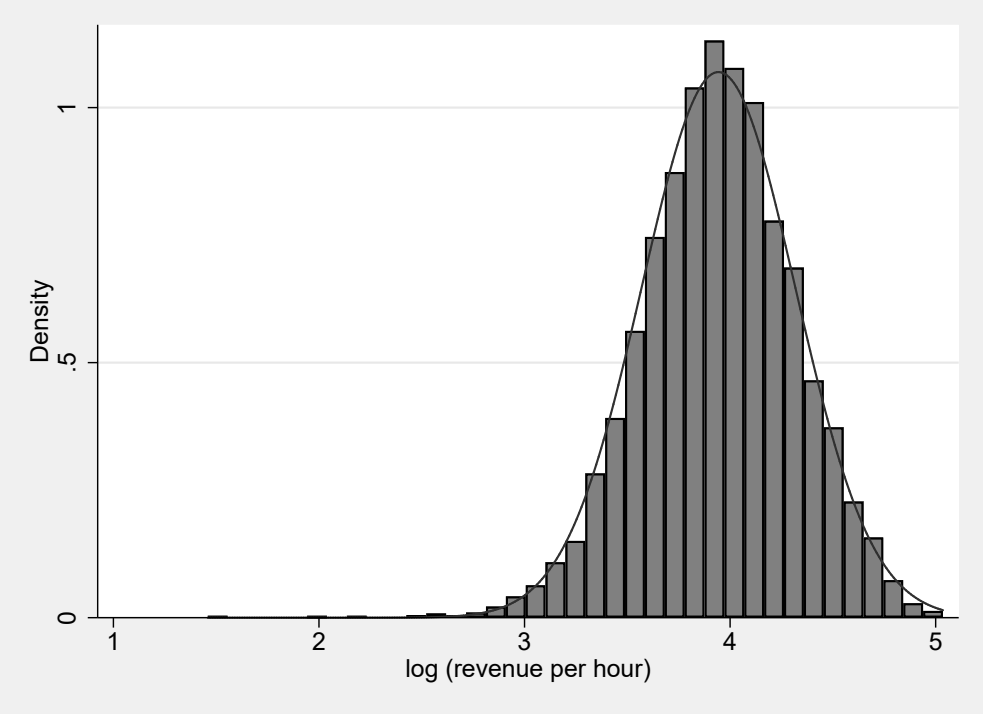
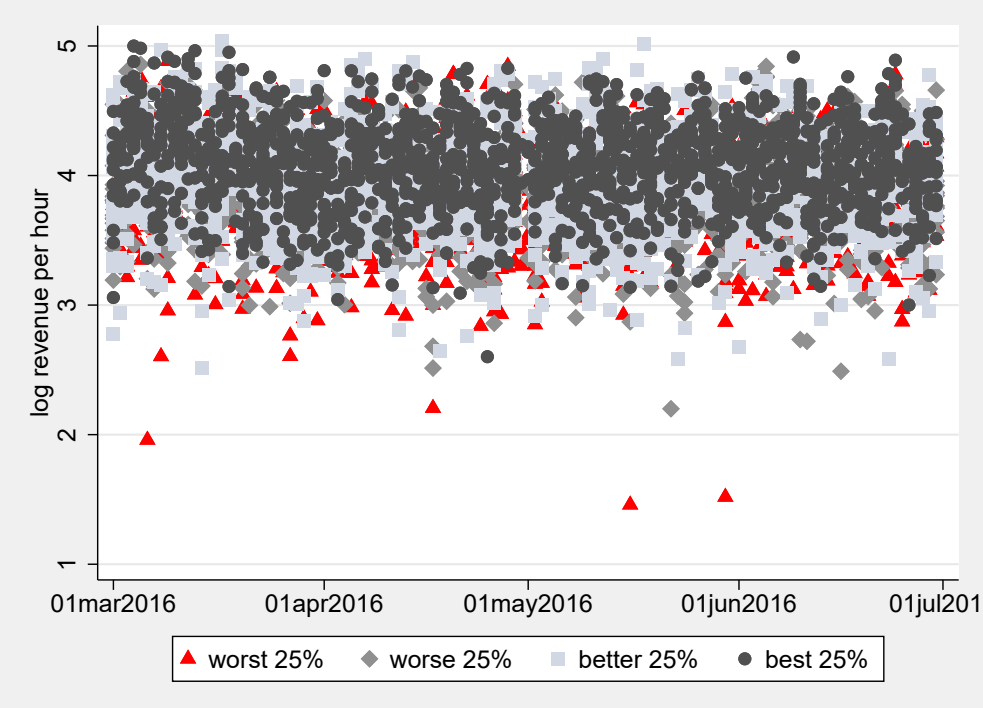


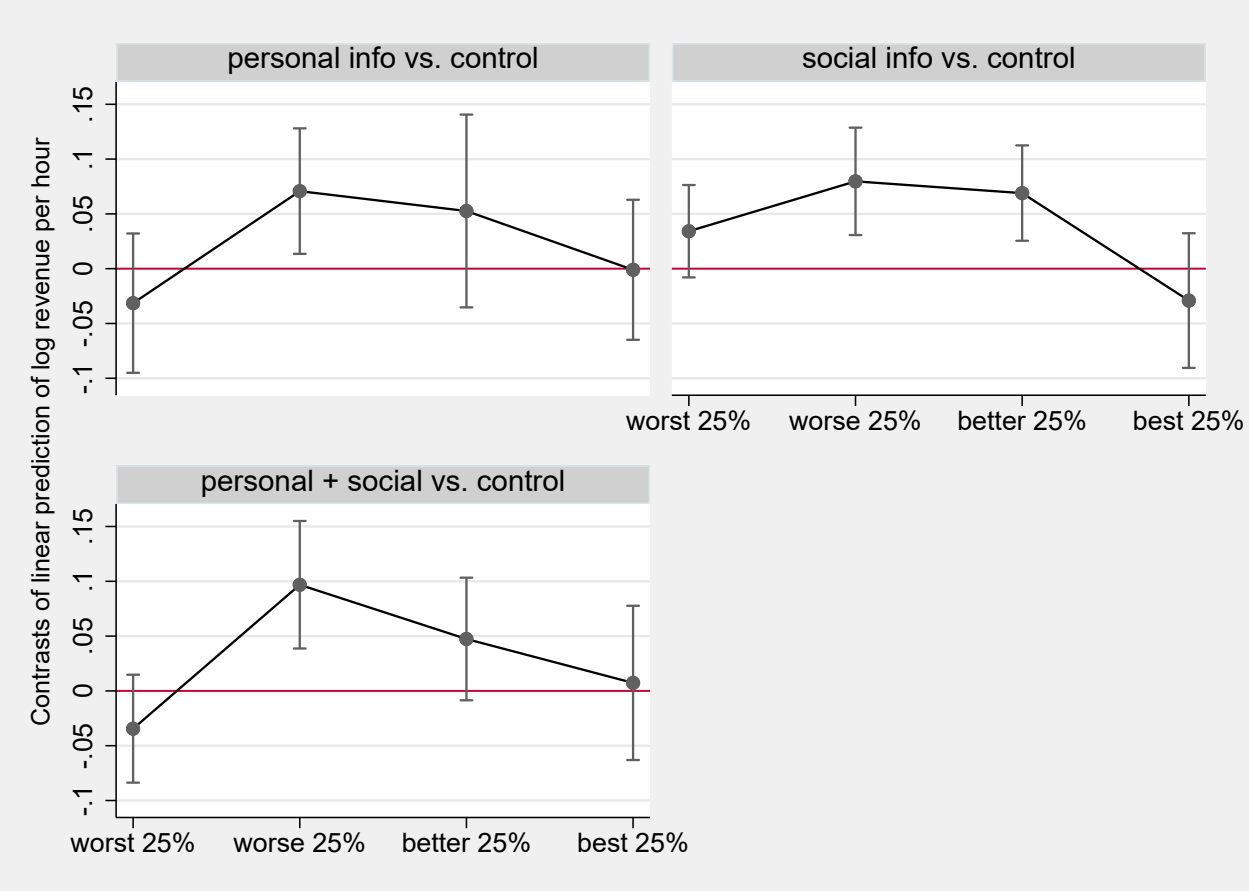
Figure 5: Histogram of the log revenue per hour in the study period



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

Appendix B Robustness checks



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 2 with fixed effects for stewards. The error bars report the 95% confidence intervals.

Figure 7: Contrasts of predictive margins for the difference-in-difference estimates

Table 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 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 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 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix C Feedback effects over time

Table 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 for more details. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 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 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 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 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 8 provides an illustration.

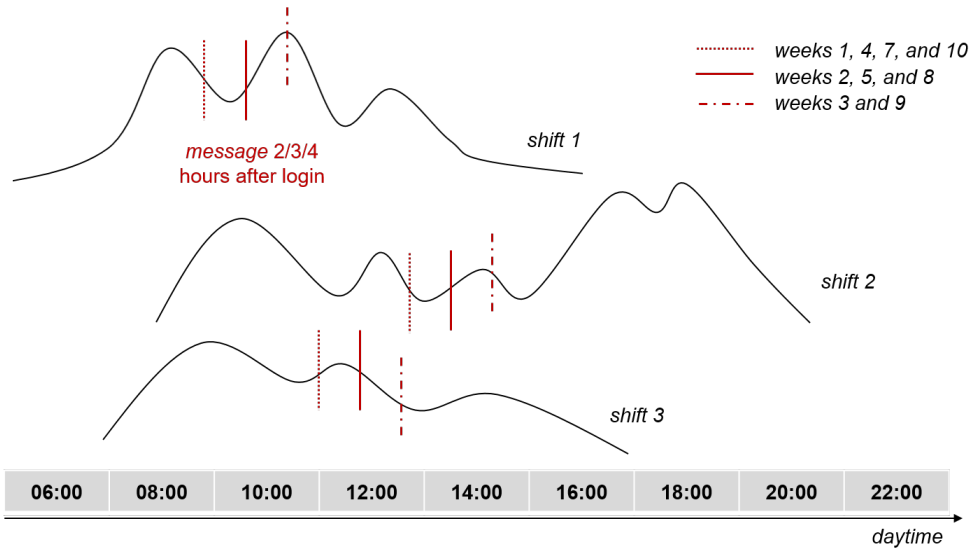


Figure 8: Illustration of the during-work feedback message

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 9 illustrates the within-treatment comparisons of the "message received" (grey) and "message not yet received" (cross-hatched) conditions. 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).

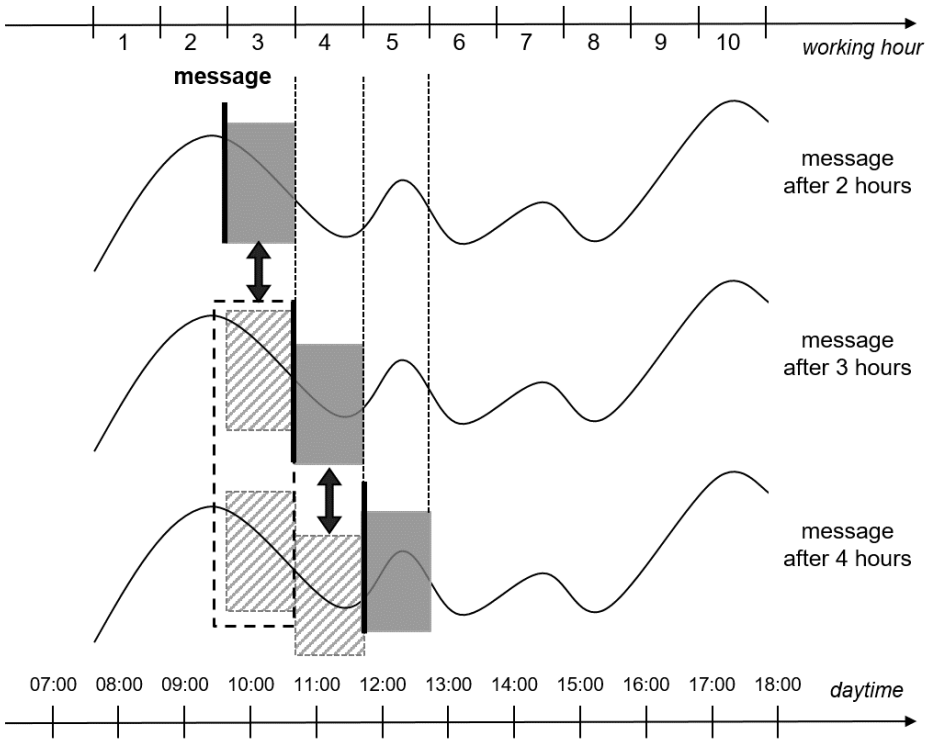


Figure 9: Model illustration

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

[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.

hour t is estimated with the following Poisson regression model:

$$y_{i,j,t} = \beta_0 + \beta_1 message_{i,j,t} + \beta_2 working\ hour\ 4_t + \beta_3 break\ time_{j,t} + \beta_4 occupancy_{j,t} + \beta daytime'_{j,t} + \beta day\ of\ week'_t j + \beta month'_j + \delta_i + \nu_j + \epsilon_{i,j,t}. \quad (3)$$

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 ($break\ time_{j,t}$) 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 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%.

D.3 Results

Table 11 shows the estimates for different specifications of Model 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 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.

[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.

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 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 10).

Notes: Poisson regression estimates (Model 3) for the number of items sold in working hours three and four. The error bars report the 95% confidence intervals.

Figure 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.

Table 12 shows the estimates of Model 3, including the interaction with a steward's current performance. *rev to go* indicates the remaining revenue per hour before steward

[20] While we were looking at the interaction with a steward's general ability in Section 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 4.1).

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.

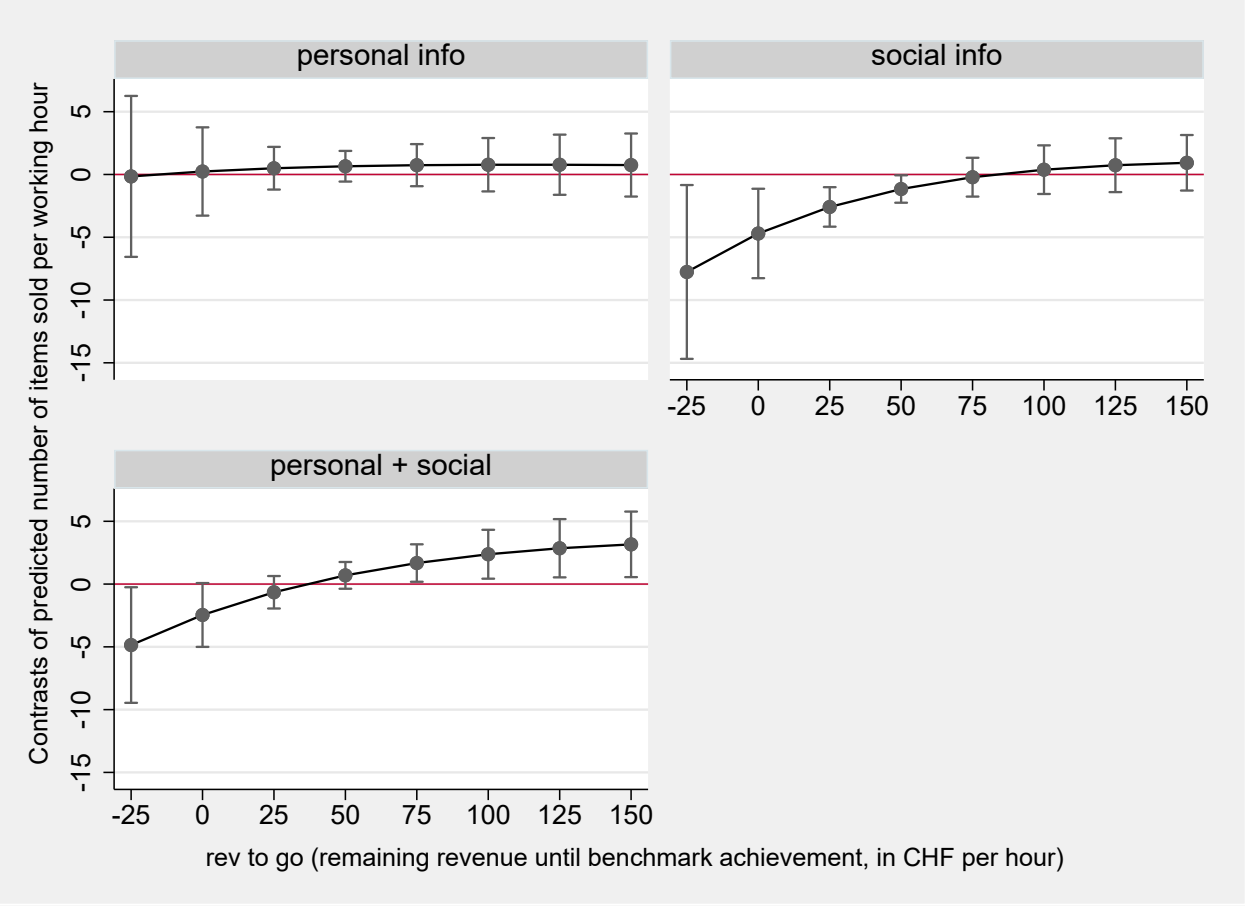
Table 12: Poisson regression: Interaction effect with current performance

	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 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 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 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 11: Contrasts of predictive margins for the feedback–performance interaction

Table 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 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 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 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.