University of Groningen

Feedback Design and Preference Elicitation: Field Experiments in Digital Economics

Romensen, Gert-Jan

DOI:
[10.33612/diss.135380025](https://doi.org/10.33612/diss.135380025)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Romensen, G.-J. (2020). *Feedback Design and Preference Elicitation: Field Experiments in Digital Economics*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen, SOM research school. <https://doi.org/10.33612/diss.135380025>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Feedback Design and Preference Elicitation: Field Experiments in Digital Economics

Gert-Jan Romensen

Publisher: University of Groningen
Groningen, The Netherlands

Printed by: Ipskamp Printing
Enschede, The Netherlands

© Gert-Jan Romensen

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted in any form or by any means, electronic, mechanical, now known or hereafter invented, including photocopying or recording, without prior written permission from the copyright owner.

Front cover: the image “Pretending to Fly” is made by Peter Nijenhuis near Kardinge, Groningen, The Netherlands, and is licensed under CC BY-ND 2.0.

university of
 groningen

Feedback Design and Preference Elicitation: Field Experiments in Digital Economics

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on
Monday 19 October 2020 at 12:45 hours

by

Gert-Jan Romensen

born on 31 January 1990
in Emmen

Supervisors

Prof. A.R. Soetevent

Prof. M.A. Haan

Assessment Committee

Prof. A.J. Dur

Prof. B.W. Lensink

Prof. J. Meer

Preface

Writing a PhD thesis is setting your first steps as a researcher, and along the way you pick up a few valuable lessons about humility, perseverance, and being grateful for the learning opportunities and the support from so many people. Colleagues, friends and family helped me to navigate through the highs and lows of the PhD trajectory. A sincere thank you is in place.

I would like to start by expressing my gratitude to Adriaan Soetevent for supervising my thesis and for being a mentor and frankly, a role model. While I can easily fill this space with praise for Adriaan's expertise and research capabilities, from which I benefited tremendously, I prefer to write instead about how much I admire Adriaan's approach to PhD supervision. From tirelessly providing feedback and career guidance to initiating numerous coffee breaks and progress meetings, Adriaan showed in every way possible that he genuinely cares about me. The work environment that Adriaan stands for is one in which I felt comfortable to express myself freely and to reach for my full potential. I have learned from Adriaan not only how to conduct research but also what a good mentor should look like. I aspire that one day I can be such a mentor to my own protégés. I will not forget the discussions about work and life in general, whether it was at the train station in Heerenveen or while walking along the IJssel in Zwolle.

As a PhD student, I was incredibly lucky to have Marco Haan as the second supervisor of my thesis. Marco's ability to quickly get to the essence of research projects really helped me in writing and positioning the chapters in this thesis. Also, without the support and flexibility of Marco it would have been impossible to implement the field experiment with the *MicroApp*, which led to Chapter 4 in this thesis. Marco also taught me to have fun in

both research and teaching. His sense of humor and taste in music made the PhD a much more pleasurable experience. As I am writing this, the Micro I playlist is playing in the background. Who knew that the Spice Girls can teach us a thing or two about consumer behavior?

I am also grateful to the reading committee, consisting of Robert Dur, Robert Lensink and Jonathan Meer. Thank you for taking the time to read my thesis and for providing valuable comments. I look up to each of you and I am really proud that you were willing to read and assess my work.

Successful implementations of field experiments often crucially depend on the tireless efforts of a select few people. I was fortunate enough to meet and work with a number of very dedicated persons, without whom the studies in this thesis could not have been implemented. At Arriva, I am especially grateful to Peter Boersma, Marten Feenstra and Wouter van der Meer for their time and support in enabling the research in Chapter 2 and Chapter 3 of this thesis. The field experiment with the MicroApp in Chapter 4 benefited significantly from the efforts of Wim-Paul Hiddink, Mark van Oldeniel and Jan Riezebos. The study in Chapter 5 that was implemented at the filling stations of TanQyou was only possible thanks to the continuing support from Jan Harmen Akkerman, Wybe Buising, Hidde van der Maas and Wesley Tankink. Thank you all for your involvement.

I also would like to thank the department of Economics, Econometrics and Finance (EEF) for providing an excellent working environment. I particularly enjoyed the many interesting conversations with my officemates: Matthijs Katz, Carolina Laureti, Ulrich Schneider, Daniël Vullings and Pieter IJtsma. Thank you also to Tadas Bružikas, Nicolás Durán, Joëlle van Essen, Tom de Greef, Pim Heijnen, Bert Kramer, Mart van Megen, Noémi Péter, Beatriz Rodríguez Sánchez, Bert Schoonbeek, Sándor Sóvágó, Lennart Stangenberg, Nannette Stoffers, Eduard Suari Andreu, Kimberley Vudinh and Doede Wiersma. I am also grateful to the people at SOM for their support: Rina Koning, Ellen Nienhuis and Kristian Peters. Thank you all for making it worthwhile to come to the Zernike Campus.

Attending conferences and presenting my research was something I enjoyed during the PhD. Meeting colleagues with similar research interests and listening to inspiring keynote speakers helped me to stay motivated.

I have particularly fond memories of the following conferences: Advances in Field Experiments (Chicago), BEER Winter School (Champéry), Ca' Foscari-Groningen PhD workshop (Venice), EEA (Lisbon), ESA (Dijon), IAREP-SABE (Dublin), IMEBESS (Barcelona) and NCBEE (Kiel). Thank you to the participants for the excellent feedback on my research.

There are countless other people that have supported me over the years. While I cannot mention them all here, I do wish to express my gratitude to some of them: Jeroen van den Born, Tom Breteler, Rick Homan, Omiros Kouvavas, Elmer Pals, Jos Postma and Sepehr Vahed. Special thanks go to my two paranymphs: Federico (Fred) Giesenow and my brother Tonny Romensen. Fred and I followed the same path in Groningen: from the honours college days all the way to the research master and the PhD. I fondly remember the many coffee breaks at the kitchenette of the ninth floor and the midnight Skype calls to discuss problem sets (“You have a cat? That’s amazing!”). I also enjoyed the many beers we had together, with the occasional plates of bitterballen. In so many ways Tonny contributed to this PhD thesis. For example, without Tonny there would be no MicroApp and the field experiment in Chapter 4 would have never been implemented. I cherish the many moments we worked together on the MicroApp, especially those when my early-morning sessions overlapped with Tonny’s late-night sessions. Thank you Tonny for being such a great brother!

Also a big thank you to the *mango’s*! Markedijk is for me synonymous with fun evenings playing Dalmuti, enjoying Vivino-approved wines, and being full for three weeks from the delicious roasted wild hare that is served during Christmas. Thank you Helma and Matt for allowing this Drent to feel at home in Twente. Phil and Annick, thank you for the many coffee breaks and for hosting Anouk when I need to do my after-dinner naps. Also, Phil, may your dream of beating me in deadlifting one day come true. Maurice and Anna, the little one is blessed with such loving parents!

My parents, John and Hilda, deserve a special mention in these acknowledgements. I still remember the moment my parents had to occupy a caravan at a caravan fair to make phone calls with teachers about my lack of motivation at high school. Who would have thought that their son would continue to write a PhD thesis? Thank you for creating a loving

environment in which I always felt supported and cared for. It is hard to describe how proud I am to be your son, but let me try to phrase it like this: if one day I am blessed to start my own family, then I cannot think of better parental role models than my own parents. I am also really grateful to my grandparents and the rest of the family for their love and support.

I would like to conclude with an apology and a thank you to Anouk, the love of my life. I apologize for the 5:00 AM alarm clock and for being glued to the laptop during many weekends. I also admit that statistical software issues are not the best conversation starter during Sunday morning breakfasts, and that the division of household chores somehow always ended up in my favor. Thank you for never complaining. My most cherished moments in life thus far I have been fortunate enough to experience together with you, from spotting a leopard in the Okavango Delta to being on the hunt for geisha in Kyoto. Honestly, though, I just as much enjoy the little moments with you: watching yet another episode of *Ik Vertrek*, gaining weight from your delicious apple pies, and not being able to keep up with you on your e-bike when visiting Drents-Friese Wold, Kardinge and Schiermonnikoog. You encourage me to become the best version of myself and to strive for the best version of us. Even though you are rightfully nicknamed “co-driver from hell” (No, I am not blind and I do see the slow-driving car that is one kilometer ahead of us), I am glad you are next to me and I am excited to explore the road ahead of us!

Gert-Jan Romensen

Groningen, 17 September 2020

Contents

1	Introduction	1
2	Worker productivity and tailored performance feedback: Field experimental evidence from bus drivers	17
2.1	Introduction	18
2.2	Related literature	21
2.3	Field setting	25
2.3.1	Industry	25
2.3.2	Research setting	26
2.3.3	Scope for improvement	27
2.4	Experimental design	30
2.4.1	Time path	30
2.4.2	Targeted peer-comparison treatments	32
2.4.3	In-person coaching	35
2.4.4	Data collection and sample construction	37
2.5	Results	39
2.5.1	Feedback reports	39
2.5.2	In-person coaching	44
2.5.3	Treatment interaction	51
2.6	Discussion and conclusion	54
2.A	Appendix	59
3	Learning from high-achieving peers: A quasi-experiment in the field	97
3.1	Introduction	98
3.2	The setting	102
3.3	The in-person coaching program	104

3.4	Data and phase-in of the coaching program	104
3.5	Results	109
3.6	Conclusion	115
3.A	Appendix	117
4	Rabbits and study habits:	
	A field experiment on pacesetters and student effort	135
4.1	Introduction	136
4.2	Related literature	139
4.3	Research question and hypotheses	142
4.4	Experimental design	143
4.4.1	Institutional setting	143
4.4.2	Phase-in of the MicroApp	144
4.4.3	Experimental conditions	146
4.4.4	Randomization and exclusion rules	148
4.4.5	Implementation plan and timeline of the experiment	152
4.4.6	Data collection	155
4.5	Empirical evaluation	157
4.5.1	First stage: no experimental variation	157
4.5.2	Second stage: experimental variation	165
4.5.3	Additional analyses	186
4.6	Discussion	188
4.7	Conclusion	190
4.A	Appendix	193
5	P(l)aying at the pump:	
	Incentivized measurement of risk attitudes in the field	217
5.1	Introduction	218
5.2	Related literature	221
5.3	Research design	223
5.3.1	The bomb risk elicitation task (BRET)	224
5.3.2	Research objectives	225
5.3.3	Field setting	225
5.3.4	Implementation plan and design choices	226
5.3.5	Timeline of the study	229
5.3.6	Data description	231
5.4	Results	233

5.4.1	Main results	233
5.4.2	Additional analyses and robustness checks	238
5.5	Conclusion	242
5.A	Appendix	243
Bibliography		255
6	Samenvatting (Dutch summary)	273

Chapter 1

Introduction

1.1 Digital technology and economic behavior

The digitalization of economic and social activities is one of the most salient and far-reaching societal developments of our time. Much of what we do and choose can be stored by digital technologies in bits and transmitted as data at much lower costs than ever before (Goldfarb and Tucker 2019). Individual behaviors that were previously hard to measure in the field, such as worker productivity and student effort, can now be minutely logged by monitoring technologies, online learning platforms, and the like. As digital technologies continue to evolve and become more ingrained in our everyday environments, policy makers and researchers have raised the question how these technologies can be used to improve economic and social outcomes.

This thesis aims to contribute to this discussion with four field experiments that explore the potential of digital technology in improving the outcomes of workers, students, and consumers. The first two experiments focus on the workplace. Productivity data from a new monitoring technology are used to detect worker-level areas of improvement, to tailor feedback, and to evaluate the impact of the feedback programs. The third experiment is in education. A digital learning platform is developed to measure student effort and to target this effort directly as a means to improve learning outcomes. The final experiment uses a communication technology in a natural field setting with many consumers to implement an incentive-compatible

experimental method that elicits individual risk attitudes repeatedly.

This section proceeds with a discussion of the research background and motivations against which the field experiments in this thesis are conducted. The discussion does not attempt to be exhaustive, but rather aims to provide a general overview of the points of departure for each experiment. Section 1.2 contains a detailed description of the chapters in which the experiments are discussed. Section 1.3 offers concluding thoughts.

Digital economics as a research field is concerned with whether and how digital technology changes economic activity (Goldfarb and Tucker 2019). Digital technology enables information about one’s behavior and choices to be stored in bits, which substantially reduces the cost of storage, computation, and transmission of data. Goldfarb and Tucker (2019) identify five types of costs that have fallen significantly in this regard: search costs, replication costs, transportation costs, tracking costs, and verification costs. This thesis zooms in on two types in particular: tracking and replication. Tracking is the ability to follow anyone’s individual behavior.¹ Replication means that data, once digital, can be replicated at near zero marginal cost.

The ability to track individual behavior likely reshapes principal-agent relationships in many domains. In traditional offline settings, an agent’s effort is oftentimes unobservable, and in such cases the principal may be forced to design a contract based on aggregate information about the agent’s activities (Hölmstrom and Milgrom 1991). In digitalized environments, an agent’s effort may be completely observable to the principal. While this increases the principal’s scope for using pay-for-performance schemes to achieve the desired effort level of the agent, such schemes are often not feasible in unionized establishments or in settings where the culture is such that financial incentives are seen as undesirable. Alternatively, the principal can use the data on observable effort to design and evaluate non-financial incentives that can motivate the agent, such as tailored performance feedback. The data can also be used to acquire a better understanding of agents’ preferences, choices, and actions. This thesis considers three principal-agent settings in more detail and explores the potential of

¹Needless to say, privacy must be warranted when working with data on individual behavior. All studies in this thesis are compliant with privacy regulations (GDPR).

digital technologies for feedback design and preference elicitation.

Worker behavior

In the workplace, digital monitoring technologies reduce the cost of tracking individual productivity and can help to personalize performance feedback (Staats, Dai, Hofmann, and Milkman 2017). Effective feedback provision to workers is important to enhance performance, especially in work environments with no pay-for-performance scheme in place (Gosnell, List, and Metcalfe 2020, Blader, Gartenberg, and Prat 2020). Earlier research has shown that management practices matter for productivity (Bloom and van Reenen 2007, Bloom, Eifert, Mahajan, McKenzie, and Roberts 2013, Syverson 2011). Now that firms increasingly have data on multiple dimensions of worker productivity at their disposal, the question that emerges is how these data should be employed in tailoring feedback.

The field experiment in Chapter 2 is motivated by the mixed evidence on the effectiveness of relative performance feedback in the workplace. While some studies show positive effects of providing rank information (Blanes i Vidal and Nossol 2011, Song, Tucker, Murrell, and Vinson 2018), other research reports that outcomes worsen after relative performance feedback (Bandiera, Barankay, and Rasul 2013, Barankay 2012). Kuhnen and Ty-mula (2012) suggest that it may be promising to customize relative performance feedback by tailoring the content or by targeting subsets of workers. The point of departure in the field experiment is the observation that existing studies on relative performance feedback report rankings on the final outcome rather than on the intermediate steps that lead to this outcome, which may be demotivating as it gives little guidance on where to improve. The experiment examines instead a relative performance feedback program in which rank information is provided on disaggregate productivity measures rather than on the final outcome, with experimental variation in the nature and number of peer-comparison messages to allow for targeting.

Chapter 2 also reports on a quasi-experimental evaluation of another approach to delivering tailored performance feedback in the workplace: in-person coaching by experienced peers. While early studies on social inter-

actions in the workplace stress that worker productivity improves mainly through peer pressure (Mas and Moretti 2009, Falk and Ichino 2006), more recent work emphasizes skills transfers and knowledge flows between workers through peer-based learning (Sandvik, Saouma, Seegert, and Stanton 2020, Lindquist, Sauermann, and Zenou 2017, Chan, Li, and Pierce 2014). Estimating the causal effect of peer-based learning on worker productivity is, however, challenging because of a previous lack of data on worker-level productivity and the exact timing and identification of peer interactions. The quasi-experiment considers a peer-based coaching program initiated by the field partner and uses coach diaries to identify worker-specific coaching moments. Productivity data from a monitoring technology are used to identify the causal effect of peer coaching in a quasi-experimental setting that exploits the phase-in design of the coaching program.

Lower replication costs induced by the use of digital technology make it significantly easier for researchers to replicate their own or others' work. Several recent studies have called for more replications in economics as a means to validate the reliability of scientific findings (Mueller-Langer, Fecher, Harhoff, and Wagner 2019, Coffman, Niederle, and Wilson 2017, Hamermesh 2007). In response to this call, Chapter 3 performs a *scientific replication* of the main findings from the quasi-experimental evaluation of the peer-based coaching program in Chapter 2. Following the definition in Hamermesh (2007), a scientific replication is replicating research findings with a different sample and a similar but possibly not identical model. The replication in Chapter 3 makes use of the fact that the exact same coaching program and monitoring technology are implemented at a different time in an area outside of the sample area discussed in Chapter 2. In this new area there is no parallel experiment in place with relative performance feedback, allowing for a test of whether the initial results extend to such a setting.

Student behavior

In recent years, there has been much interest in the use of digital technologies in education (Escueta, Nickow, Oreopoulos, and Quan forthcoming). Existing research in this area has mainly focused on comparing courses with

digital learning components to more traditional courses with offline learning components (Swoboda and Feiler 2016, Brown and Liedholm 2002). Less research is available that examines the potential of educational technologies in measuring and stimulating student effort to improve learning outcomes. In light of concerns that student effort is declining (Oreopoulos, Patterson, Petronijevic, and Pope 2018, Babcock and Marks 2011), and that this may have repercussions for academic performance (Metcalf, Burgess, and Proud 2019, Stinebrickner and Stinebrickner 2008), there is an urgent need for more research on how students can be motivated to study more.

In most of the existing research on ways of increasing student effort, actual effort is unobserved due to data limitations. Output measures (exam scores) are used instead to approximate the provided effort, but these measures are likely confounded by ability. Disentangling the effects of effort and ability is important in the design of educational interventions (Levitt, List, Neckermann, and Sadoff 2016). Being able to target student effort directly may be more promising when the objective is to motivate students to study (Clark, Gill, Prowse, and Rush 2017, Lavecchia, Liu, and Oreopoulos 2016). Designing policies that do so is a challenge, however, as it requires data on effort and a way of dealing with commonly observed self-control problems among students (Augenblick, Niederle, and Sprenger 2015, Wong 2008).

In chapter 4, a field experiment is presented in which an online learning platform is developed and introduced in a large university course to measure and target student effort directly. The experiment starts with the observation that many educational settings have long put in place commitment mechanisms to mitigate self-control problems among students, such as intermediate deadlines on the completion of a task. Evidence on the effectiveness of such mechanisms is mixed (Koch, Nafziger, and Nielsen 2015). The typical behavioral pattern that is observed is that of bunching of effort just before the deadline. To counteract such behavior, recent studies propose goal bracketing as a commitment mechanism, whereby students endogenously set subgoals and adopt mental accounts to evaluate their performance in a broad (e.g., weekly) or narrow (e.g., daily) account (Koch and Nafziger 2017, Koch and Nafziger 2016, Hsiaw 2015). Narrow goals can be particularly motivating for students with self-control problems because

it rules out behavior in which lack of effort today can be compensated for with more effort in the future, though crucially at the expense of flexibility.

In the field experiment, a novel digital commitment mechanism is developed and tested that is fully consistent with students' initial study plans. The mechanism, inspired by pacesetters in sports, incorporates the motivating effect of goal bracketing in a dynamic fashion and preserves flexibility. The online learning platform is used to present the mechanism to students in real time, and to experimentally evaluate its effects on study effort and learning outcomes through A/B-testing. This way, valuable information is obtained not only on the realizations of individual student effort, but also on how these realizations compare to the initial study plans of the students. Escueta et al. (forthcoming) conclude that using educational technology to tailor and administer behavioral interventions is a promising way forward. The field experiment is one of the first attempts in this direction.

Consumer behavior

In business-to-consumer (B2C) communication, firms more often use digital technologies to interact with consumers in once traditional offline settings. This opens up opportunities for the on-the-fly measurement of consumer behavior and preferences in everyday decision-making environments. It likely also changes the trade-offs that researchers consider when they select methods for preference elicitation in the field. Consider, for example, risk preferences. Individual risk attitudes are central in many decisions involving uncertainty, and its measurement has therefore attracted much attention from researchers. A challenge identified in the literature is the trade-off between the practical feasibility of eliciting the risk attitudes of large samples in the field and making sure that the elicited attitudes are incentive compatible (Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner 2011).

Two approaches to eliciting risk attitudes have become popular, each emphasizing a different side of the trade-off. Survey-based approaches are fast and relatively inexpensive to implement, but generally do not have a built-in incentive-compatibility mechanism. Experimental approaches are incentive compatible, but they are not often employed in large field studies

due to time and cost considerations, complexity, and infrastructure needs.

Chapter 5 shows that the widespread adoption of digital technologies in B2C communication importantly increases the scope for the incentive-compatible measurement of individual risk attitudes in the field. The chapter reports on a field study in collaboration with a startup in the Dutch retail gasoline industry that has an on-site B2C communication technology. The technology is used to repeatedly administer both incentivized (Bomb Risk Elicitation Task) and non-incentivized (SOEP general risk question) risk elicitation methods to a large group of consumers in a natural field setting, with the aim of comparing the qualitative estimates of both methods.

The technology enables the tracking of individual responses through time in both the BRET (Crosetto and Filippin 2013) and in the SOEP general risk question (Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner 2011), and as such may offer new insights on the stability of risk attitudes in field settings (Barseghyan, Molinari, O'Donoghue, and Teitelbaum 2018).

1.2 The chapters in this thesis

Chapter 2. Worker productivity and tailored performance feedback: Field experimental evidence from bus drivers

In Chapter 2, we collaborate with a large public transport company to investigate how digital technologies can be used in the workplace to design and evaluate tailored performance feedback. During the study period, the company phased in electronic on-board recorders (EOBRs) in its bus fleet in several concession areas in the Netherlands. These EOBRs meticulously measure multiple dimensions of bus driver performance and other trip-related variables, thereby providing an unusually rich and detailed picture of worker productivity through time. Next to fuel economy, the EOBR system records per driver the trip-level performance on acceleration, braking, and cornering (ABC). The output thus contains both aggregate and disaggregate measures of worker performance, and hence creates a particular setting that is ideal for studying questions pertaining to the design and impact evaluation of tailored performance feedback programs.

We use the EOBR data from one concession area with 409 tenured bus drivers to examine the potential of two forms of tailored feedback to workers: written peer-comparison feedback and in-person coaching. The previous section highlighted that studies on relative performance feedback typically report feedback on the final outcome rather than on the intermediate steps that lead to this outcome, perhaps due to data limitations. As a result, feedback recipients may be left with little or no guidance on which dimensions to improve to ensure better future outcomes. The field experiment in this chapter builds on the idea that one's level of control over intermediate steps is usually higher than over final outcomes, as with students deciding how many hours to study for an upcoming exam (see also Chapter 4 in this thesis). A higher level of control may improve motivation.

Our approach is the following. We first create peer groups with drivers from the same base location, and then use the monthly relative performance of a driver in this group to construct peer-comparison messages that notify the driver whether he or she is doing well or poorly on one or more of the ABC dimensions of driving behavior (disaggregate productivity measures). The messages are integrated in monthly feedback reports and are tailored in the sense that there is experimental variation in the nature (corrective or positive messages) and composition (only corrective messages or a mix of corrective and positive messages) of the peer-comparison messages. This way, we can examine how firms can use monitoring technologies, like the EOBR system, to target feedback messages and to direct the attention of workers to individual areas of improvement.

Using over 500,000 trip-level observations in a two-year sample period, we find that the targeted peer-comparison feedback generally does not improve the aggregate (fuel economy) and disaggregate (ABC) measures of driving behavior when compared to a control group in which no such feedback is distributed. Only for the subset of drivers not exposed to prior in-person coaching there is suggestive evidence that driving behavior improves after receiving corrective feedback messages. One possible explanation is that once drivers have met a coach who gave them detailed feedback on what they do right and wrong on a trip, they become insensitive to subsequent messages about their relative performance.

The in-person coaching program is implemented in parallel to the peer-comparison feedback in the said concession area and aims to facilitate knowledge flows and skills transfers between peers. In a coaching session, an experienced peer rides along with a bus driver as a designated coach for a portion of the driver's shift, and provides tailored feedback afterwards based on observed actions (coaches have no access to the EOBR data).

Chapter 2 presents a quasi-experiment in which we aim to identify the short-run and long-run impact of peer-based coaching on individual driving behavior. We exploit the fact that all coaches minutely maintained diaries in which they logged all coaching activities. These logs pinpoint for all drivers in the sample when exactly they were exposed to coaching. After presenting evidence that the timing of coaching can be considered the outcome of a quasi-random phase-in assignment, we employ a differences-in-differences approach to estimate the impact of participating in a single on-the-road coaching session. The EOBR data make it possible to not only identify the immediate impact, but also to trace the effects over time.

The results indicate that coaching leads to significant improvements on all dimensions of driving behavior for drivers in the bottom half of the performance distribution. The improvements last for about eight weeks, which suggests that drivers do not habituate to the more economical driving style. Further actions may therefore be needed that can counter the decay effects and better enable drivers to make economical driving a lasting habit.

Existing studies on peer effects in the workplace often document that social pressure serves as the main channel through which peer effects operate. In these studies, however, workers perform an easy task with few learning opportunities, and can readily observe coworkers' productivity to infer an acceptable pace of working. Such job features are increasingly absent in many modern workplace settings, like the public transport sector discussed in Chapter 2. While peer pressure may partially account for improved driving performance during the coaching session, it cannot explain why improvements persist in the ensuing weeks. The results are therefore more in line with the interpretation that workers, at least temporarily, acquire new skills and knowledge through learning from peers.

Chapter 3. Learning from high-achieving peers: A quasi-experiment in the field

One possible concern of the research setting in Chapter 2 is that two tailored feedback programs are implemented and studied in parallel, and that there may be interactions between the two programs in ways that are not fully identified by the presented analyses. To completely rule out that such interactions affect the qualitative results of the peer-based coaching program, I evaluate the impact of the same coaching program in two other concession areas of the public transport company with no tailored peer-comparison feedback in place.

Chapter 3 presents the results from this evaluation. As in Chapter 2, I use detailed coach logs to identify all driver-specific coaching moments in the two-and-a-half year sample period. A similar quasi-experimental research design is adopted that makes use of the phase-in feature of the coaching program. Using close to two million trip-level observations from 1,032 tenured bus drivers, I obtain results that are qualitatively very similar to those reported in Chapter 2. Bus drivers significantly improve their driving behavior on multiple dimensions after participating in a coaching session, but the effects generally wane after a few weeks. Interestingly, the pattern of decay in treatment effects is less pronounced for acceleration and braking, suggesting some habit formation in these dimensions. Drivers in the bottom half of the performance distribution improve the most again after receiving feedback from an experienced peer.

The results from Chapters 2 and 3 combined indicate that in-person coaching by experienced peers is a promising way of facilitating knowledge and skills transfers between workers. Especially workers with low initial productivity stand to benefit the most from receiving tailored feedback from high-achieving coworkers. The observed decay in treatment effects suggests that additional steps are needed to sustain improvements. One promising approach to explore in future research is to use productivity data from monitoring technologies to detect when exactly decay in performance sets in for a particular worker. Coaches can then be assigned to these workers for more tailored repeat sessions that focus on the specific issues at hand.

Chapter 4. Rabbits and study habits: A field experiment on pacesetters and student effort

In Chapter 4, I develop and use a new educational technology to implement a field experiment in a large university course with 573 students. Existing research indicates that many students may perform below their academic potential due to a simple lack of study time. Recent studies also suggest that motivating students to study may be better achieved when effort is incentivized rather than output, but only limited progress on this is made thus far due to a lack of data on effort and students' initial study plans. Procrastination and self-control problems further exacerbate the challenge of getting students to study more.

In the experiment, I elicit detailed weekly study goals and plans from all students and use these to construct individualized pacesetters (*rabbits*). Pacesetters are moving reference points that visualize the preferred study pace of the present self by moving exactly according to the initial study plan, akin to pacesetters in marathons. Falling behind the pacesetter immediately confronts the student with his or her procrastination tendencies. Under the assumption that students exhibit reference-dependent preferences, loss aversion makes falling behind psychologically painful and motivates the student to catch up and exert more effort.

The field experiment is designed to identify the causal effect of the pacesetter on student effort and learning outcomes, and to control for confounding effects related to the acts of formulating study goals and plans. To measure student effort and to administer the pacesetter intervention in real time, I develop a new educational technology that is offered free-of-charge as a non-mandatory study tool in a large course at a Dutch university. The technology is an online learning platform that contains extensive self-developed practice material and recaps. Combining platform usage data with data from weekly in-class surveys on time management, I am able to map precisely how student-level effort is distributed through time over multiple learning inputs. Administrative databases with learning outcomes and student background information complement these data on study effort.

I find that the pacesetter induces students to set more ambitious study

goals, about twice as high than in the control group, but that this does not lead to more study effort exerted. The pacesetter has no impact on study time, as measured by the educational technology, nor do I observe effort spillovers to other learning inputs and other courses. Further analyses show that students already fall behind their pacesetter at the start of the week and on average show no tendency to catch up later in the week. I consider multiple measures of learning outcomes and find no effect from the pacesetter on these outcomes. Treatment heterogeneity based on gender, ability, and procrastination is considered in detail but found to be absent or limited in terms of effort, goals, and learning outcomes.

More generally, I aim to show with this field experiment the potential of using educational technologies as a means to target behaviors that have a negative impact on academic performance. I show how such technologies can be used to measure student effort in real time, to administer personalized interventions, and to implement experiments through A/B-testing.

Chapter 5. P(1)aying at the pump: Incentivized measurement of risk attitudes in the field

In Chapter 5, we collaborate with a startup in the retail gasoline industry to investigate how, in an incentive-compatible way, the risk attitudes of a large and diverse group of consumers can be elicited in natural field settings. Previously, a lack of digital infrastructures in offline field settings made it costly and complex for researchers to elicit individual risk attitudes in an incentivized way. Because of this, experimental elicitation methods have thus far mainly been confined to laboratory studies with student samples.

In this chapter, we argue and show that the widespread adoption of business-to-consumer (B2C) communication technologies increases the scope for the incentivized measurement of individual risk attitudes in the field. We pioneer the approach of digitally integrating risk-elicitation methods with other decision tasks in the field by reporting the results of a one-year study conducted at thirteen geographically-dispersed filling stations in the Netherlands. Filling stations are arguably one of the best natural settings to find a wide cross-section of the population above the legal driving age.

Using the digital communication infrastructure at these filling stations, we ask gas station visitors once a month via the pump touch screens to complete the Bomb Risk Elicitation Task (an incentive-compatible method) and to answer the SOEP general risk question (a non-incentivized survey question) immediately after they completed the process of filling the tank. In the BRET, introduced and validated by Crosetto and Filippin (2013), subjects need to decide how many boxes to collect out of a set of 100, one of which contains a bomb. Earnings are linearly increasing with the number of boxes collected, but are zero if the box with the bomb is among the collected boxes. The SOEP risk question, validated by Dohmen et al. (2011), asks subjects to rate their general willingness to take risk on a scale from 0 (unwilling to take risks) to 10 (very willing to take risks). The infrastructure at the filling stations makes it possible to administer these elicitation methods in a timely manner and to uniquely identify all subject choices through time.

In the one-year sample period, about 2,500 visitors of the filling stations completed the two elicitation methods. We find that the majority of subjects in this field setting, about 82%, submit choices in the BRET that are consistent with risk aversion. We observe no gender differences in BRET choices, nor do we observe that other demographic factors, such as age and proxies for wealth, can explain some of the variance in the number of boxes collected in the BRET. The modal response to the SOEP question (seven) suggests that subjects rate themselves as moderately willing to take risks. This is particularly the case for males. The correlation between choices in the BRET and in the SOEP question is positive but small. In repeat elicitations, we observe that the elicited risk attitudes are time-varying.

Firms and other organizations increasingly use digital channels to communicate with their clients. These digital infrastructures offer novel opportunities for eliciting risk attitudes directly in the environment in which actual financial decisions are made. In doing so, the predictive power of the elicited risk attitudes with respect to these financial decisions may improve. It may also offer new opportunities for organizations and researchers to tailor products and services based on consumer risk profiles and needs.

1.3 Overview and concluding thoughts

To conclude, this thesis presents the results from field experiments that explore the potential of digital technologies to measure economic activity, to tailor interventions, and to improve outcomes. Chapters 2 and 3 focus on the potential in the workplace by showing how the technology-enabled measurement of individual worker productivity offers new opportunities for companies to personalize and evaluate performance feedback programs. Chapter 4 shows how digital technology can be used in education to measure and target student effort directly as a means to improve learning outcomes. Chapter 5 illustrates how technologies can be used in the field to interact with consumers in real time and to elicit individual preferences on the fly.

The chapters show in three different settings how digital technology has given principals the means and data to steer and learn about behavior and preferences of agents. This way, the thesis contributes to the literature on feedback design and preference elicitation. Chapter 2 shows that tailoring peer-comparison feedback to workers' individual areas of improvement is generally ineffective in improving worker productivity. Workers do not perform better when they are told on which disaggregate productivity dimensions they perform well or poor relative to peers. Chapters 2 and 3 provide evidence that being coached by experienced peers leads to significant, albeit temporary, improvements on multiple productivity dimensions, especially among those workers with poor initial productivity. The evidence is such that the gains from coaching are due to actual learning and not due to social pressure. Chapter 4 shows that students do not study more when they are presented a real-time pacesetter that is based on their own study plans. Even though students actively engage with their pacesetter, by setting non-zero study goals that are more ambitious than those set by control students, there is no evidence that the pacesetter leads to improved learning outcomes. Chapter 5 reports that more than 80% of the subjects in a natural field setting with a large non-student population are risk averse. In this setting, the correlation between choices in an incentivized and a non-incentivized risk elicitation method is found to be positive, but small in magnitude. Age and gender were generally not predictive of risk attitudes.

Given the scale and pace at which digital technology is transforming the lives of many, there seems to be no shortage of interesting research questions that can be asked. This thesis identifies at least three fruitful avenues for further research. First, digital technology has dramatically increased the scope for tailoring, targeting, and timing behavioral interventions based on rich data of individual behavior. Second, technology-enabled just-in-time measurement of individual preferences in the field offers many opportunities for near real-time adaptations of products and services to personal needs. Third, more research is needed on the relative strengths of humans vis-à-vis technology in realizing knowledge flows and skills transfers in organizations.

Chapters 2 and 4 demonstrate the potential of digital technology in the design of behavioral interventions, but also showed that it is not a panacea for fostering behavioral change. As organizations have access to increasingly rich profiles of individuals, I envision that in influencing people's choices, personalized behavioral interventions will gain much more traction at the expense of the more traditional use of monetary incentives. In online markets, low tracking costs did not lead to personalized pricing, but instead led to personalized advertising (Goldfarb and Tucker 2019).

Chapter 5 shows how digital technology enables just-in-time elicitation of individual preferences in the decision-making environment of consumers. As researchers continue to discuss the stability (Schildberg-Hörisch 2018) and domain specificity (Einav, Finkelstein, Pascu, and Cullen 2012) of risk preferences, having access to just-in-time elicited measures may help researchers to better understand the consumer choice under consideration. Products and services may be adapted on the fly based on these measures.

A more general research theme that emerges from this thesis is the question what can best be done by digital technology and what not. While the studies in this thesis show that digital technology is very capable of measuring behavior and identifying individual areas of improvement, it may not be the best communicator of the feedback that follows from this. Human interactions may still be essential in realizing knowledge flows and skills transfers in the workplace, in the classroom, and in other settings. More research on this theme is much needed, as it is ultimately about our unique strengths as human beings in a digitalized world.

Chapter 2

Worker productivity and tailored performance feedback: Field experimental evidence from bus drivers*

Abstract

How should performance feedback be tailored to improve worker productivity? In a natural field experiment with 409 bus drivers and over 500,000 trip-level observations, we test the potential of two forms of individual feedback: written peer-comparison feedback and in-person coaching. We find that the announcement of the written feedback program has a substantial and significant effect on fuel economy and outcomes pertaining to passenger comfort; targeted peer-comparison feedback is generally ineffective; in-person coaching generates significant improvements on all dimensions for drivers in the bottom half of the performance distribution for about eight weeks; in-person coaching reduces the impact of written peer-comparison feedback but not vice versa.

*This chapter is based on Romensen and Soetevent (2020a).

2.1 Introduction

Giving effective performance feedback is critical in maintaining and enhancing worker productivity, especially in work environments that hinder the use of pay-for-performance schemes (Blader, Gartenberg, and Prat 2020, Gosnell, List, and Metcalfe 2020). The adoption of digital monitoring technologies in the workplace has greatly expanded managers' scope for giving workers tailored feedback (Staats, Dai, Hofmann, and Milkman 2017), and offers researchers the opportunity to study the effects of changing personnel practices on within-firm productivity (Bartel, Ichniowski, and Shaw 2004).¹ As these technologies routinely record data on the underlying dimensions of worker-level productivity, it is now possible for managers and researchers to target directly the behaviors that give rise to aggregate productivity. How can performance feedback best be tailored to target these behaviors?

This chapter contributes to answering this question by collaborating with a large public transport company that is in the process of installing electronic on-board recorders (EOBRs) in its entire bus fleet. For every worker, EOBRs measure the trip-level aggregate productivity (fuel efficiency) and the underlying performance on acceleration, braking, and cornering (ABC).² We use this setting to evaluate two promising forms of tailored performance feedback: targeted peer-comparison feedback and in-person coaching. The two forms are tested using over 500,000 trip-level observations from 409 tenured bus drivers in a two-year sample period.

Targeted peer-comparison feedback is evaluated using a natural field experiment. The experiment is motivated by the heterogeneity in findings related to the effects of relative performance feedback in the workplace (see Section 2.2 for a discussion of the literature). We reason that part of this heterogeneity may be due to studies typically reporting feedback on final outcomes rather than on the behaviors that lead to these outcomes. Rank information may in such cases be demotivating as it gives little guidance

¹Recent studies that examine how the adoption of electronic monitoring technologies by firms has impacted worker productivity include Pierce, Snow and McAfee (2015) and Kelley, Lane and Schönholzer (2018).

²See Baker and Hubbard (2003) and Hubbard (2003) for early work incorporating the EOBR technology. They study how the adoption of on-board computers has influenced the decision of truckers to integrate or outsource trucking services.

on where to improve, possibly leading to feelings of not being in control. Workers may be more motivated by relative rankings if they are reported on productivity dimensions that are specific and under direct control.

The experiment is designed to test this. We use the EOBR data to determine for every driver his or her relative performance on each of the ABC dimensions compared to a group of similar peers. This information is used to identify individual areas of improvement and to construct peer-comparison feedback messages on these dimensions. The messages are integrated in monthly individual feedback reports that are newly introduced as part of the company's EcoManager campaign on economical driving. To examine how firms can use monitoring data to target individual areas of improvement, we introduce experimental variation in the dosage and composition (corrective or positive) of the peer-comparison messages.

Specifically, we randomly allocate drivers into four groups. In the control group, drivers only receive general feedback in the report and no peer-comparison messages. In the second group, drivers receive a maximum of one corrective message on a dimension that requires attention, allowing for the focusing of effort. The third group is similar, except that negative feedback is supplemented with positive feedback in case a driver who performs poorly on some dimensions scores well on others. This allows us to assess the value of providing a mix of corrective and positive feedback. Drivers in the fourth group receive corrective messages on all dimensions if performance is poor across the board. Together, these interventions enable us to explore the potential of digital monitoring technologies in customizing relative performance feedback, such that it enhances worker motivation.

In-person coaching by experienced peers is another promising way of tailoring feedback to disaggregate productivity measures. Recent studies have stressed the importance of peer-based learning for knowledge flows and skills transfers between workers (Sandvik, Saouma, Seegert, and Stanton 2020, Lindquist, Sauermann, and Zenou 2017, Chan, Li, and Pierce 2014). Using a quasi-experimental design, we evaluate the effects of an in-person coaching program that was implemented in parallel to the targeted peer-comparison feedback. In this program, designated experienced drivers engage in coaching their colleagues by riding along with a bus driver for a

portion of the driver's shift. At the end of the ride, the coach evaluates the trip in detail and gives tailored tips for improvement. We use detailed coach diaries to pinpoint all driver-specific coaching moments in the data.

Due to the hop-on hop-off approach to coaching and regulations that disallow coaches access to the driver's performance, the timing of coaching can be considered the outcome of quasi-random assignment: coaches select the drivers they will coach on a given day in a way that is unrelated to a driver's past performance. Our empirical evidence corroborates this. We exploit this phase-in feature of the program to uncover the short-run and long-run causal impact of a coaching session on driving behavior.

In a meta-analysis of feedback interventions, Kluger and Denisi (1996) conclude that many studies lack a control group against which the observed effects can be contrasted. In such studies, feedback eligibility and intensity are likely to correlate with workers' (relative) productivity outcomes. This sample selection biases estimates of feedback effectiveness that are based on comparisons of worker productivity just before and right after the worker has received feedback. The (quasi-)random assignment of drivers to the feedback designs considered in this chapter avoids these selection problems.

Our main findings are as follows. First, the launch of the EcoManager campaign reduces fuel consumption by 0.4 liters/100km (0.40 SD, $p < 0.001$). Distributing the feedback reports generates a further 0.1 SD reduction ($p = 0.006$). For the targeted peer-comparison feedback, we find zero effects. Varying the number and nature of peer-comparison feedback messages has no additional impact on worker productivity.

Second, we observe strong and immediate effects of coaching. On the day of coaching the fuel need reduces by 0.6 liters/100km (0.58 SD, $p < 0.001$) and the number of acceleration events by 1.1 events/10km (0.50 SD, $p < 0.001$). For braking and cornering behavior, these effects are less pronounced and not (braking) or less (cornering) significant. The improvements due to coaching tend to persist with a smaller magnitude in the ensuing weeks but fade out after about seven to nine weeks. Zooming in, we find the impact of coaching on performance confined to drivers in the bottom half of the performance distribution.

Third, we find a nonreciprocal relation between in-person coaching

and targeted peer-comparison feedback: prior exposure to peer-comparison messages does not change the effectiveness of in-person coaching for any of the productivity measures. Peer-comparison feedback, however, is only effective in the group of drivers that did not yet receive in-person coaching. One possible explanation is that once drivers have met a coach who gave them feedback on what they do right and wrong on a trip, they become insensitive to subsequent written messages about their relative performance.

Fourth, in the group of non-coached drivers, those in the treatment with the maximum number of negative messages and no positive comments show the largest improvement in productivity outcomes. In other words, limiting negative feedback or mixing negative with positive feedback does not seem to have any beneficial effect. This shows that it is important to pay attention to interactions between the different elements of job design.

This chapter proceeds as follows. Section 2.2 reviews the related literature. Section 2.3 describes the field setting of the study. Section 2.4 elaborates on the research design, provides further details on both feedback programs and presents the data. The empirical analysis of both programs follows in Section 2.5. Section 2.6 discusses the results and concludes.

2.2 Related literature

A large literature shows that management practices matter for worker productivity (Bloom and van Reenen 2007, Bloom, Eifert, Mahajan, McKenzie, and Roberts 2013, Syverson 2011). Despite a considerable body of empirical work on relative performance feedback, however, the question how rank information affects worker productivity has not yet received its definite answer. Previous studies indicate that relative performance feedback can improve worker productivity (Blanes i Vidal and Nossol 2011, Song, Tucker, Murrell, and Vinson 2018), sales growth (Delfgaauw, Dur, Sol, and Verbeke 2013) and (high school) student performance (Tran and Zeckhauser 2012, Azmat and Iriberry 2010). Other studies report decreased performance following the provision of rank information (Ashraf, Bandiera, and Lee 2014, Bandiera, Barankay, and Rasul 2013) and improved performance when this information is abolished (Barankay 2012).

People may exhibit rank incentives (Barankay 2012, Tran and Zeckhauser 2012) when relative performance information affects self-image (Benabou and Tirole 2006) and status (Moldovanu, Sela, and Shi 2007). These rank incentives can lead to demotivation at the bottom of the performance distribution, which reduces the average effects of feedback programs that rely on social comparisons (Ashraf, Bandiera, and Lee 2014). Kuhnén and Tymula (2012) suggest that it may be promising to customize relative performance feedback by tailoring the content or by targeting subsets of workers. Blader, Gartenberg and Prat (2020) for example find that the provision of relative performance information in plants with(out) a teamwork culture leads to decreased (improved) truck driver performance.

What may account for some of the heterogeneity in results is that rankings are typically reported on final outcomes rather than on the intermediate steps leading to these outcomes. In this form, the message may be demotivating because it gives little guidance on where to improve and signals that improvement requires one big step rather than several small and clear steps. Feedback provision on disaggregated productivity measures can provide much more guidance on where to improve, making it easier for workers to change their behavior. It may empower poor performers by increasing the feeling of control, raising awareness of behaviors that require attention, and by offering suggestions for specific actions that workers can take. The feeling of being in control is a key source of human motivation (Ryan and Deci 2000).

Our research design does exactly that. One possible concern with disaggregated relative performance feedback, however, is that it may aggravate the adverse effects of feedback provision. That is, it may make poor performance even more salient to workers at the bottom of the distribution. When information directly enters the utility function (Golman, Hagmann, and Loewenstein 2017), informing workers about poor performance on multiple dimensions may decrease motivation.³ Also, the increased level of detail in the written feedback may trigger adverse effects similar to those caused by feedback overload. Increasing the feedback frequency can lead

³For example, there is evidence that reward-related brain areas correlate negatively with lower relative incomes (Dohmen, Falk, Fließbach, Sunde, and Weber 2011).

to more mistakes (Eriksson, Poulsen, and Villevall 2009) and reduced task effort due to overwhelmed cognitive resources (Lam, DeRue, Karam, and Hollenbeck 2011). This poses a challenge, as poor performers have the biggest room for improvement and are thus precisely the group that one wishes to target with detailed feedback. Should feedback thus be delivered on all dimensions simultaneously to prevent drivers from underperforming in non-reported dimensions (Hölmstrom and Milgrom 1991, Baker 1992) or should feedback be limited to prevent information overload (Simon 1973, Hitt and Brynjolfsson 1997, Edmunds and Morris 2000)?

Treatment effect heterogeneity may also show in the drivers' response to in-person coaching. A prevalent finding in the literature on peer effects in educational outcomes (Sacerdote 2011) is that high-ability students benefit most from the presence of high-ability peers (Fruehwirth, 2013, Hoxby and Weingarth, 2005, Lavy, Paserman, and Schlosser, 2011, Lavy, Silva and Weinhardt, 2012), although some studies (Burke and Sass 2013) find that students with the lowest past performance gain most from exposure to higher-achieving peers.⁴ Drivers in our design are coached by experienced colleagues assigned the role of coach. Hence, a coaching session explicitly exposes a driver to a high-achieving peer.

While recognizing the differences between a school environment and the work environment that we study – both in the nature of the interactions and the outcomes of interest – the cited studies on peer effects suggest that the effect of in-person coaching may depend on a driver's own past performance. Our study checks whether this result on peer effects carries over to non-educational contexts. A related study is Sandvik, Saouma, Seegert and Stanton (2020) who run a field experiment among salespeople. They similarly find that exposure to a high-achieving peer generates productivity gains, but in their setting the gains persist even after twenty weeks. Chan, Li, and Pierce (2014) also observe lasting effects from peer-based learning. They conclude that peer-based learning is more important for worker productivity growth than learning-by-doing.

⁴Booij, Leuven and Oosterbeek (2017) find that low-ability students benefit from having low-ability peers but that high-ability students are unaffected by their peer group composition.

Next to contributing to the empirical literature on optimal feedback design in operations management, our findings also address the broader societal challenge of how to combat unsustainable energy consumption practices. While there has been much progress in our understanding of non-financial incentives in residential energy consumption, research on how these insights generalize to firms is scant (Gerarden, Newell, and Stavins 2017, Gosnell, List, and Metcalfe 2020, Nilekani 2018).⁵ Our work aims to partly fill this gap and should be viewed as part of the emerging literature that looks at the workplace for evidence on the effect of non-financial incentives on conservation efforts (Gosnell, List, and Metcalfe 2020).

Given that firms increasingly record and store data on multiple dimensions of worker-level productivity, tailoring feedback by decomposing consumption into its underlying sources seems a viable and promising approach to creating novel data-driven designs of conservation incentives (Brynjolfsson and McElheran 2016). The setting of a transport company is apt as the transport sector takes a heavy toll on the environment, accounting for one-fifth of global primary energy use and one-quarter of energy-related carbon dioxide (CO₂) emissions (IEA 2012). Indeed, the International Council on Clean Transportation hails fuel-efficient driving as low-hanging fruit to improve conservation levels (ICCT 2013).⁶ We examine how this can be achieved in a setting where drivers have no financial stake in fuel savings.

⁵Existing studies on non-financial incentive schemes in the residential sector stress the importance of feedback and social approval in increasing welfare (Allcott and Mullainathan 2010). For example, incorporating social comparisons in feedback reports reduces household consumption of energy (Allcott 2011, Ayres, Rasemand, and Shih 2013) and water (Ferraro and Price 2013), with long-run effectiveness depending on whether households alter their capital stock of habits or physical technologies (Allcott and Rogers 2014). Recent research, however, also notes that social comparisons can trigger asymmetric effects (Holladay, LaRiviere, Novgorodsky, and Price 2016) and may interact with other non-financial incentives when stimulating green behavior (Hahn, Metcalfe, Novgorodsky, and Price 2016). This has reinforced the need for detailed evaluations of non-financial incentives pertaining to energy efficiency and also raises the question how these findings generalize to workers. Allcott and Kessler (2019) emphasize the importance of incorporating the (moral and emotional) costs incurred by nudge recipients in assessing the welfare effects of social comparisons.

⁶Barkenbus (2010) has sketched the potential of multidimensional eco-driving campaigns and feedback mechanisms for personal transportation. We instead examine the extent to which this potential can be realized in public transportation.

2.3 Field setting

2.3.1 Industry

Our field partner is Arriva, a European-wide passenger transport company operating various transport modes in public transport. Bus transport is the firm's largest business unit.⁷ In the Netherlands, bus concessions are granted to companies by means of a tendering procedure.⁸ Winning a tender gives companies the exclusive rights to operate in a designated area for a number of years. To stimulate firms to engage in environmentally friendly behavior and to improve the living conditions of its citizens, local governments let environmental objectives feature prominently in the requirements tendering parties need to meet.⁹ This has geared public transport companies toward the use of environmentally friendly technologies.¹⁰

In the long run, this trend may drive bus companies to buy vehicles with a hybrid or electric fuel technology. On a shorter time horizon, the installment of electronic on-board recorders (EOBRs) helps the companies to meticulously measure performance on several dimensions of driving behavior. For example, the version used by Arriva records trip-level performance on fuel consumption and comfort dimensions such as acceleration, braking and cornering (ABC). Each driver logs into the system with a unique personnel number to match the performance records and trip-related background variables. This enables precise monitoring and provides managers and researchers with a wealth of high-frequency data on worker productivity and conservation efforts.

The system works as follows for the comfort dimensions. Based on test rides under different circumstances, threshold performance levels are formulated by the company for every dimension. Technically, the thresholds

⁷At the time of the study, Arriva Group is part of Deutsche Bahn, employs over 60,000 people and annually delivers more than 2.2 billion passenger journeys in 14 European countries.

⁸See the Passenger Transport Act 2000.

⁹Interested companies are commonly requested to submit a sustainability plan in which they indicate how they decrease the ecological footprint of public transport in the concession area.

¹⁰The Dutch Ministry of Infrastructure mentions public transport as a "trend setter" in the area of sustainable technologies (MIVW, 2010, p. 87).

relate to minimum G-force measurements by a three-axis accelerometer in the bus. During each trip, the EOBR records an ‘event’ whenever an action by the driver is in excess of these thresholds. The performance measure of the ABC dimensions is the number of events per 10km, with fewer events indicating better driving behavior. The outcome data can subsequently be linked with centralized databases containing information on a host of driver and trip characteristics. This allows us to get a detailed picture of driver performance over time under various on-the-road conditions.

2.3.2 Research setting

As part of its EcoManager campaign, Arriva Netherlands installed new EOBRs in its entire fleet in the time period 2015-2017. The EOBR data will be used as input to monthly feedback reports that will be distributed among the drivers. In addition, a new coaching program is introduced in which drivers receive real-time feedback and advice from an experienced colleague during on-the-road sessions. The new technology and the feedback programs are phased in over time in the concession areas.

We join the implementation process in the first concession area, comprising about two-thirds of a province in the Netherlands and serving about 5.16 million travelers in a year.¹¹ The majority of drivers in this area are tenured employees, while a small number (about 14%) operates on a temporary contract. Most of the drivers are experienced and have a long career of driving buses or other vehicles. They are typically not involved in other tasks within the organization. Opportunities for promotion are limited and the work council is against using financial incentives to reward good performances.¹² In the past, drivers received no personal feedback.

Each driver belongs to one of the six base locations (usually a municipality) in the area and operates on routes that are stipulated by the

¹¹Based on the official number of electronic check-ins with the public transport card in 2015.

¹²Within firms, the design of conservation incentives is often dictated by institutional constraints that hinder the use of pay-for-performance schemes. See e.g. Freeman (1981), who finds that within-establishment dispersion of wages is narrower in unionized establishments. He attributes this in large part to unions’ wage practices, such as the adoption of uniform wages (rather than merit-based pay).

concession. For five locations, virtually all routes are between cities and in rural areas. Routes are based on timetables and do not vary much over time. One location (the largest one) has a mixture of urban and rural routes. Urban trips are mostly operated by a special bus type that runs on natural gas. Within a location, drivers' weekly shifts rotate. This implies that the worker faces week-to-week variation in his or her assignment to trips and the schedule repeats after about 14 weeks. This way of scheduling ensures that drivers are familiar with their routes and drive each route under different on-the-road circumstances. The schedules provide ample within-location variation in the type of trips, such that all drivers face a more or less similar mixture of relatively easy and difficult trips. Because of the rotation of shifts, multiple drivers are assigned a given route. Taken together, this variation allows us to include a rich set of fixed effects in our empirical analysis.

2.3.3 Scope for improvement

Before discussing the research design, we wish to get an idea of the potential scope for improvement by considering the factors that influence driver performance on fuel economy and the ABC dimensions. What part of performance can be influenced by the driver and what part is caused by external factors such as weather and traffic conditions? For fuel economy, we observe sizable between-driver variation in performance. To drive 100km, the average driver uses 24.91 liters of fuel, with a standard deviation of $\sigma = 2.30$.¹³ Table 2.1 shows that part of this variation can be attributed to differences in driving conditions.

The first column shows that the bus type accounts for 27.9 percent of the between-trip variation in fuel economy, with the Intouro and longer buses having a sizable and significantly worse fuel economy. The impact of weather conditions (column (2)) seems limited. Fuel economy is – as one expects – negatively correlated with the number of stops per kilometer, the

¹³25 liters/100km $\sim$ 10.6 gallon/100miles. Throughout the text, we will state (changes in) fuel economy in l/km instead of km/l because of the miles-per-gallon (MPG) illusion (Larrick and Soll 2008). Figure 2.A.1 shows the entire distribution of driver fixed effects for the outcome variable fuel economy.

Table 2.1: Determinants of Fuel Economy

Dependent variable:	Fuel economy				
	(1)	(2)	(3)	(4)	(5)
VDL 10m	-1.013*** (0.029)	-1.014*** (0.029)	-0.948*** (0.028)	-0.943*** (0.027)	-0.945*** (0.016)
VDL 14m	8.910*** (0.165)	8.958*** (0.158)	8.904*** (0.145)	8.935*** (0.146)	8.922*** (0.136)
Intouro	4.785*** (0.239)	4.808*** (0.239)	4.445*** (0.234)	4.487*** (0.216)	4.483*** (0.161)
Rush hour 7-10am			-0.349*** (0.031)	-0.350*** (0.030)	-0.248*** (0.015)
Rush hour 4-7pm			0.334*** (0.039)	0.334*** (0.038)	0.246*** (0.018)
Non-scheduled trip			0.095 (0.093)	0.077 (0.091)	0.096** (0.044)
No. of stops per km.			0.721*** (0.087)	0.750*** (0.086)	0.715*** (0.083)
Trip length (in km.)			-0.023*** (0.001)	-0.025*** (0.001)	-0.024*** (0.001)
Ln(No. of passengers)			1.263*** (0.016)	1.288*** (0.016)	1.275*** (0.014)
Punctuality			0.013*** (0.005)	0.010** (0.005)	0.045*** (0.003)
Constant	23.716*** (0.053)	23.450*** (0.057)	20.531*** (0.134)	21.274*** (0.182)	21.156*** (0.166)
R ²	.279	.291	.409	.43	.514
Number of trip-level observations	533171	533171	533171	533171	533171
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Fuel economy in liters/100km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Trip punctuality is the difference in minutes between actual and scheduled driving time. Weather data are collected from a weather station located in the regional capital and are maintained by the Royal Netherlands Meteorological Institute (KNMI). Standard errors are clustered by driver.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.2: Driver Variation in Fuel Economy and ABC Comfort Outcomes

Driving dimension	μ	σ_r	c_v	$\Delta(p_{90} - p_{10})$	Residual correlation		
					Acc.	Braking	Corn.
Fuel economy	24.91	1.03	0.04	2.46	0.399	0.083	0.166
Acceleration	10.87	2.22	0.20	4.87		0.268	0.271
Braking	1.68	1.02	0.61	0.93			0.154
Cornering	2.27	1.93	0.85	1.71			

Notes: Fuel economy: liters/100km. ABC dimensions: no. of events/10km. σ_r : residual standard deviation; c_v : coefficient of variation. $\Delta(p_{90} - p_{10})$: difference in performance 90th vs. 10th percentile driver.

number of passengers, evening rush hours and the bus running late. These variables seem to capture most of the day-to-day variation in fuel economy, as adding day fixed effects only slightly improves the R^2 . Structural differences in driver performance explain an additional eight percentage points of variation in trip-level fuel economy (column (5)). When we control for the rich set of trip characteristics as in column (5) of Table 2.1, the variation in performance between drivers as measured by the residual standard deviation is $\sigma_r = 1.03$.¹⁴ Hence, the potential for improvement is economically significant: A policy able to move a driver's average fuel economy from the 90th percentile to the 10th percentile reduces this driver's fuel bill by 2.46 liters/100km or about 10%.

We use the residual variation σ_r to compute the coefficient of variation $c_v = \sigma_r/\mu$ as a standardized measure of dispersion to compare the relative scope for improvements in fuel economy and in the ABC dimensions. For fuel economy, this coefficient equals 0.04 ($= 1.03/24.91$). The numbers for the ABC dimensions are shown in Table 2.2. The coefficients of variation show that in relative terms, between-driver dispersion is larger for the ABC dimensions than for fuel economy. However, for braking and cornering the average number of events per 10km is relatively close to the absolute lower bound of zero, thereby limiting the upward potential for a large fraction of drivers.

Of course, the different outcomes are related: more acceleration events for instance increase fuel consumption. Table 2.2 shows the residual correlation between the fuel economy and comfort dimensions after controlling for the same set of trip-level characteristics as in column (4) of Table 2.1. Fuel economy is correlated with acceleration and, to a lesser extent, with cornering. This supports the focus on the ABC dimensions in our peer-comparison treatments. Next to being worker productivity measures in their own right, improvements in either of them also contribute to fuel economy.

¹⁴Appendix section 2.A.2 provides details on the estimation of σ_r , and section 2.A.3 contains the corresponding tables for the ABC dimensions.

2.4 Experimental design

2.4.1 Time path

Figure 2.1 depicts the timeline of the study. First, we use the old on-board system to establish a long baseline of fuel consumption, starting in January 2015. At this stage, drivers are not informed about the upcoming feedback, nor that they are being monitored. The new EOBR system enables the collection of comfort dimensions baseline data in the months September and October 2015. The company sent promotion material about the EcoManager campaign to the different locations on October 5, 2015. The project was officially launched with a kickoff event on November 9, 2015. At this date, the LED-array in the buses is also switched on, providing drivers with some instant feedback.¹⁵ At the event, all drivers were informed about the digital monitoring and the introduction of monthly individualized feedback reports starting in December 2015.

Figure 2.1: Timeline Study

Peer-comparison feedback The second period (Nov. 9-Dec. 15, 2015) is used to disentangle effects of the announcement and LED activation

¹⁵The LED-array contains eight LEDs: three green, two amber and three red. The green LEDs illuminate when the driver is in the ‘sweet spot zone’, determined by the (vehicle dependent) rotations per minute of the engine. The LEDs indicate the occurrence of an ABC event by flashing three times one second. As these events can only be timed when an action by a driver exceeds the threshold, any LED-array indication happens ex post.

from the feedback effect. In the third period (Dec. 15, 2015-Nov. 15, 2016) drivers receive their monthly feedback reports with peer-comparison feedback. Finally, the post-experimental period (Nov. 15, 2016 - Jan. 31, 2017) starts with a one-time notification to the drivers that the peer-comparison messages are no longer included in the reports.¹⁶

Previous research has shown that workers adjust their effort in response to a feedback announcement, even though they have not yet learned any new information from the first feedback round (Blanes i Vidal and Nossol 2011). The company's decision to separate these events is a convenient feature of our research setting. Drivers were informed during the announcement period that the feedback will not be used in formal evaluations. This may rule out career concerns as an alternative explanation, but note that it runs counter to the firm's objectives to follow through on this claim (Hölmstrom 1979).

Apart from the feedback programs under consideration, no other incentives were used by the company to promote conservation efforts among workers. In the spirit of Barankay (2012), the one-time notification message is included at the end of the experiment in order to examine the effect of a withdrawal of peer-comparison messages.

In-person coaching The face-to-face coaching program that runs in parallel starts around the kickoff event in November 2015. Most drivers receive their first coaching in the weeks following the kickoff event.¹⁷ During this period, the company reserved extra time for the coaches to ride along with drivers and to answer questions related to the upcoming feedback. Coaching intensity gradually decreases until it levels off after the first feedback report in mid December 2015. In a few cases, drivers participated in additional coaching sessions (55 drivers, 18% of all coached drivers). We control for these additional sessions in our analyses. We have complete coach logs

¹⁶The precise text of this message is as follows (translated from Dutch): "Dear colleague, starting this month, this report will no longer include information about your performance relative to your colleagues". This message was part of the report that was distributed in November 2016 to all drivers that were part of the treatment conditions with peer-comparison feedback (all drivers except those in the control condition).

¹⁷Appendix section 2.A.4 shows the timing of coaching.

for the period till April 30, 2016. Some coaches indicated that they no longer provided or kept track of coaching after April 2016. For this reason, we restrict attention to the period till April 30, 2016, in our evaluation of the coaching program. Thirty-two drivers (10% of all coached drivers) received coaching prior to the feedback announcement.

2.4.2 Targeted peer-comparison treatments

Fueled by the conviction that the biggest gains in fuel-efficient and comfortable driving can be made when behaviors with the largest room for improvement are targeted, the company wants the peer-comparison feedback messages to emphasize the dimensions on which the driver can improve. The treatment variation in peer-comparison messages is integrated into the monthly feedback report received by all drivers.¹⁸

Drivers are randomly assigned an experimental condition, stratified along the dimensions of base location, gender, and years of service at the company. We construct reference groups in which driver performance on each comfort dimension is compared to colleagues with the same base location and treatment status.¹⁹ This creates a natural and homogeneous comparison group for drivers in which competition is likely to generate strong incentives (Lazear and Rosen 1981, Delfgaauw, Dur, Sol, and Verbeke 2013). The comfort dimensions are disaggregated measures of driving behavior over which drivers have a strong direct influence, thereby making the feedback as concrete and useful as possible to the recipients.

At the start of each month, the company shares with us a summary of each driver's performance during the previous month. We use this information to assess how a driver performed compared to his/her peers and to assign peer-comparison messages.²⁰ Dependent on treatment assignment,

¹⁸A sample feedback report can be found in Appendix section 2.A.5.

¹⁹This is because pre-treatment information revealed that high and low scores are occasionally concentrated in base locations. Limiting peer-comparison groups to drivers with the same treatment status ensures that reference groups are relatively small – such that drivers have a reasonable chance of earning (avoiding) a positive (negative) message – and avoids indirect treatment interference.

²⁰The performance summary contains information on the bus-specific percentile rank of the driver on each driving dimension (compared to all drivers in the concession area who also operated on that bus type in the previous month). The final percentile rank for

a number of negative (positive) messages are provided if a driver belongs to the bottom 50% (top 25%) of the reference group.

Treatment T1 [0n0p] is the control condition with no peer-comparison messages. In treatment T2 [1n0p], one negative message is provided if drivers underperform on a particular dimension. That is, they are explicitly informed that they rank poorly compared to peers and are encouraged to improve. In T3 [1n1p], drivers additionally have a chance of receiving one positive message. In this case, they are made aware of their good ranking and are encouraged to keep up the good work. If a driver performs poor (or well) on multiple dimensions, one will be randomly chosen. Finally, in T4 [3n0p], drivers run the risk of receiving corrective feedback on all comfort dimensions. Using T3 [1n1p] as an example, the precise (translated) text of the messages reads as follows:

Dear colleague, In terms of taking corners, you belong to the top 25 percent of the bus drivers in your location. You are doing excellent on this dimension! In terms of braking, you belong to the bottom 50 percent of the bus drivers in your location. You can improve on this dimension!

A printed version of the report is delivered around the 15th day of each feedback month via the team manager or pigeonhole. Drivers in the control condition receive the same feedback report but without the targeted messages, so as to account for general feedback effects.²¹ The report contains general feedback in the form of a letter score, ranging from A (highest score) to D (lowest score) on the comfort dimensions and fuel economy. Furthermore, it contrasts the overall score of the individual driver with the score of his or her base location. Table 2.3 summarizes the experiment.

each driving dimension is the sum of the percentile ranks of the driver on each bus type, weighted by the number of kilometers driven on that bus type in that month. Within a reference group, a driver's final percentile rank determines how (s)he has performed compared to his/her peers.

²¹Working with an uninformed control group is not possible due to company policies requiring that every driver should at least receive some feedback. By handing out reports to drivers in the control condition, we embed the experimental variation more naturally and explicitly recognize and control for Hawthorne and general feedback effects.

Table 2.3: Experimental Conditions

Conditions	General feedback	Max # positive message(s)	Max # negative message(s)
T1 [0n0p]	Yes	0	0
T2 [1n0p]	Yes	0	1
T3 [1n1p]	Yes	1	1
T4 [3n0p]	Yes	0	3

At this point, it is important to stress that the treatments condition on the eligibility to receive negative (and positive) peer-comparison feedback but not on the actual exposure. For example, among individuals in treatment group T2 [1n0p], only about 70% of the drivers receive a negative message in a given feedback round because they score lower than half of their peers on at least one of the three comfort dimensions. In case they perform poorly on multiple dimensions, one is selected randomly for peer-comparison feedback. The remaining 30% performs well on all dimensions and is therefore not notified with a message. Hence, the treatment effects that we present show the effect of treatment eligibility. They are conservative estimates of the effect of exposure to peer-comparison feedback, as only part of the group actually receives these messages in a given month. For each driving dimension, there is considerable month-to-month variation in the group of drivers in the top-25% and bottom-50% group. While most drivers move in and out, some drivers are never in the top (bottom) part.²²

Table 2.4 summarizes per experimental condition the data in the final analysis sample and reports the outcome of balance tests. The p -values show that driver pre-experimental performance in terms of the outcomes fuel economy and ABC events is well-balanced across the groups. A comparison of a rich set of trip-level and bus-type characteristics also reveals no differences across experimental groups, indicating that drivers in the different treatment groups on average have been exposed to very comparable driving conditions.²³ Drivers are on average 54 years old and work for 20 years at the company. Most drivers are male (89%). The average trip had

²²For instance, on acceleration, 19% of the treated drivers is never in the bottom-50% (and 16% always); 42% are never in the top-25% (and 9% always). Outcomes are similar for braking and cornering. Appendix section 2.A.6 gives a detailed overview per treatment condition of the number of messages sent per month and the driving dimensions targeted.

²³For each of the dimensions along which we stratified (base location, gender, years of service), $p \geq 0.99$.

a length of 31 km and was typically driven in rural areas (84%).

In sum, the detailed data allow for precise identification of good and bad performers in every feedback round. The peer-comparison messages are subsequently intended as a means to assist drivers in offering guidance on where to improve or maintain performance. They are updated every round to inform about progress and to avoid drivers from slacking off. The treatment variation enables us to vary the intensity of the corrective and positive feedback drivers receive.

2.4.3 In-person coaching

In parallel, the company initiated a coaching program. Six experienced drivers (one for each base location) were recruited as coaches based on their track record of driving behavior. All coaches participated in a training on how to approach drivers and how to communicate feedback. Since coaches are bus drivers themselves, there is only limited time available for coaching activities (about one day every two weeks).²⁴ Furthermore, because of the hop-on hop-off approach to on-the-road coaching, a coach's previous session determines the choice set for the next. This makes random allocation of coaching sessions at the driver-trip level impossible. At the same time, it is next to infeasible for coaches to target specific drivers, also because coaches have no access to the individual feedback reports and hence cannot target drivers with poor scores. We will provide empirical support for the view that the assignment of drivers to coaching is the outcome of a quasi-random process.

In a coaching session, a coach rides along with a bus driver for a portion of the driver's shift. This allows the coach to personalize the feedback and to direct attention to the driver-specific issues at hand. A session is not announced to the driver beforehand. The coach writes down examples of what goes well and wrong and identifies obstacles that may hinder driver performance, such as sharp corners. Due to the presence of passengers, there is no or limited interaction between the driver and the coach during the ride. The coach provides feedback once the trip is completed and pas-

²⁴Coaches can decide which day they use for coaching. They vary the day of the week such that every driver has a chance of being coached.

Table 2.4: Descriptive Statistics of Experimental Conditions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Full sample		T1 (0n0p)		T2 (1n0p)		T3 (1n1p)		T4 (3n0p)		Joint test: treatment effect = 0	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.		
Pre-experimental performance												
Fuel economy	25.20	(2.37)	25.28	(2.21)	25.00	(2.47)	25.28	(2.50)	25.24	(2.31)	0.33	[0.80]
Acceleration	12.30	(6.43)	12.38	(6.65)	11.91	(6.59)	12.28	(6.27)	12.61	(6.27)	0.20	[0.90]
Braking	3.00	(4.39)	3.20	(4.63)	2.98	(4.60)	2.59	(3.66)	3.25	(4.63)	0.46	[0.71]
Cornering	2.88	(4.93)	2.99	(5.11)	2.85	(5.17)	2.62	(4.60)	3.06	(4.89)	0.16	[0.93]
Demographics												
Year of birth	1962	(8.28)	1962	(8.83)	1962	(8.62)	1962	(7.62)	1962	(8.12)	0.01	[1.00]
Year of employment	1996	(11.53)	1996	(11.65)	1996	(11.48)	1996	(11.32)	1996	(11.83)	0.04	[0.99]
% share of FTE≥0.9	76.28		75.73		74.51		79.41		75.49		0.26	[0.86]
% share of female drivers	10.51		10.68		9.80		10.78		10.78		0.02	[0.99]
Trip-specific variables												
Punctuality	-2.88	(0.80)	-2.80	(0.86)	-2.94	(0.76)	-2.80	(0.86)	-2.96	(0.72)	1.13	[0.33]
Distance traveled	31.36	(13.21)	31.32	(13.99)	31.89	(12.59)	31.14	(12.84)	31.10	(13.53)	0.08	[0.97]
Number of passengers	13.26	(3.71)	13.38	(3.91)	13.27	(3.69)	13.18	(3.60)	13.23	(3.69)	0.05	[0.98]
Number of bus stops	37.84	(8.79)	37.38	(9.09)	38.54	(8.40)	37.93	(8.76)	37.51	(8.99)	0.35	[0.79]
% share of rides:												
- Morning rush hours	19.45		19.34		20.53		18.64		19.31		0.46	[0.71]
- Evening rush hours	19.53		19.99		18.20		20.55		19.39		0.72	[0.54]
- Weekend	14.09		14.31		14.01		3.98		14.03		0.03	[0.99]
- Holidays	9.99		9.88		10.27		9.71		10.10		0.27	[0.85]
- Urban area	15.84		16.77		13.68		15.89		17.00		0.19	[0.91]
- School	0.8		0.7		0.8		0.8		0.7		0.21	[0.89]
% share of trips on bus types												
Bus type VDL	75.20		73.64		77.71		75.63		73.88		0.33	[0.81]
Bus type Intouro	9.65		10.21		9.34		9.30		9.75		0.10	[0.96]
Bus type IRIS	15.15		16.15		12.96		15.07		16.37		0.21	[0.89]
Base locations (# drivers)												
Location 1	12		3		3		3		3		0.00	[1.00]
Location 2	61		16		15		15		15		0.01	[1.00]
Location 3	30		7		8		8		7		0.05	[0.99]
Location 4	74		19		18		18		19		0.02	[1.00]
Location 5	150		37		38		38		37		0.02	[1.00]
Location 6	82		21		20		20		21		0.02	[1.00]
Number of drivers	409		103		102		102		102			

Notes: Unit of observation is the driver. Columns (11) and (12) show F -statistics and p -values from a balance test of whether the treatment coefficients are jointly equal to zero. Pairwise t -tests with the control group T1 do not reveal any statistically significant differences in means for any of the variables for any of the treatment groups at the $p = 0.10$ level. Data are from EOBRs in buses and centralized databases with driver and trip characteristics. Fuel economy in liters/100km. Performance on the comfort dimensions (Acceleration, Braking and Cornering) as the number of events/10km. The pre-experimental period is the period before receiving the first feedback report. Punctuality is the difference in minutes between actual and planned driving time. Distance traveled is measured in kilometers. Number of passengers based on check-ins with public transport cards. VDL and Intouro buses have diesel engines, the IRIS bus runs on natural gas. Morning and evening rush hours are from 7:00-10:00 and 16:00-19:00, respectively. Holiday rides take place during, for example, Christmas, New Year's Eve and school holidays. School rides are along routes with schools and universities as final destinations. Stars indicate a statistically significant difference in means with the control group.

sengers have left the bus. The trip is reconstructed using the written-down examples. Both personal and general advice are offered that focus on fuel consumption, punctuality and the ABC dimensions.²⁵ Drivers are treated as equals and feedback is delivered in a constructive and positive manner.

Coaches maintain a detailed log of their activities, allowing us to pinpoint when and how often drivers are coached. We use these logs to pin down the coaching date. To check whether the assignment of coaching sessions is quasi-random and not based on pre-selected criteria, we compare for each outcome variable (fuel economy and ABC) the mean baseline performance of drivers who have received their first coaching and non-coached colleagues with the same base location. Table 2.5 verifies balancing on multiple baseline outcome performance measures and covariates. We present both the standard p -values and the ones adjusted for the problem of multiple hypothesis testing using the Bonferroni and Holm correction. Only for morning and evening rush hours we find statistically significant differences. These differences merely reflect that coaches tend to start their work early in the morning. For none of the other variables, we find differences that are even close to significance, especially once we take into account multiple hypothesis testing. This supports the view that the implementation of the coaching program exhibits a quasi-random order of phase-in.

2.4.4 Data collection and sample construction

The EOBRs are installed in all three bus types the company operates. The VDL bus is most commonly used, accounting for about 75% of all trips performed. Intouro buses are mainly used for routes with a long travel distance. Two specific features importantly distinguish the IRIS bus from the other bus types. First, both the VDL and Intouro buses have diesel engines, but the IRIS bus runs on natural gas. This implies that for trips completed with an IRIS bus no records on fuel economy are available. Second, whereas the VDL and Intouro buses are used by drivers in all base locations, the IRIS bus is only used by drivers in the largest and most urbanized base location. Hence, the treatment effects on the outcomes fuel

²⁵These notes are not included in the logs, so unfortunately we do not know exactly what the coached has conferred with a driver.

Table 2.5: Quasi-Random Coaching: Balancing Tests on Baseline Performance and Non-Performance Descriptives

	C	NC	$\Delta(\text{C-NC})$	stand. p -value	Bonf. corr.	Holm corr.
<i>Baseline performance</i>						
Fuel economy	25.038	25.247	-0.209	0.0865	1.0000	0.2595
Acceleration	13.132	13.349	-0.217	0.5200	1.0000	1.0000
Braking	3.715	3.789	-0.075	0.7744	1.0000	1.0000
Cornering	1.093	1.189	-0.096	0.2065	1.0000	1.0000
<i>Share of experimental conditions</i>						
T1 (0n0p)	0.260	0.247	0.013	0.8118	1.0000	1.0000
T2 (1n0p)	0.244	0.261	-0.016	0.7602	1.0000	1.0000
T3 (1n1p)	0.235	0.228	0.007	0.8968	1.0000	1.0000
T4 (3n0p)	0.260	0.264	-0.003	0.9532	1.0000	1.0000
<i>Demographics</i>						
Year of birth	1962.4	1961.8	0.558	0.3372	1.0000	1.0000
Year of employment	1996.2	1996.3	-0.150	0.8706	1.0000	1.0000
Share of FTE $\geq$ 0.9	0.801	0.776	0.025	0.6205	1.0000	1.0000
Share of female drivers	0.094	0.073	0.020	0.554	1.0000	1.0000
<i>Trip-specific variables</i>						
Punctuality	-2.942	-3.018	0.076	0.2251	1.0000	1.0000
Distance traveled (in km.)	30.564	32.025	-1.461	0.0996*	1.0000	0.3984
Number of passengers	15.035	15.283	-0.248	0.5983	1.0000	1.0000
Number of bus stops	37.934	37.864	0.070	0.9268	1.0000	1.0000
Share of rides:						
- Morning rush hours	0.298	0.159	0.139	0.0076***	0.1976	0.0076***
- Evening rush hours	0.125	0.253	-0.127	0.0084***	0.2184	0.0168***
- Weekend	0.031	0.031	0.000	1.0000	1.0000	1.0000
- Fill in	0.004	0.025	-0.021	0.1539	1.0000	0.7695
- Holidays	0.119	0.125	-0.006	0.8839	1.0000	1.0000
- Urban area	0.149	0.146	0.003	0.9366	1.0000	1.0000
- School	0.004	0.004	0.000	0.9954	1.0000	1.0000
<i>Share of rides on bus types</i>						
VDL	0.773	0.755	0.017	0.7427	1.0000	1.0000
Intouro	0.078	0.099	-0.021	0.5555	1.0000	1.0000
IRIS	0.149	0.146	0.003	0.9366	1.0000	1.0000

Notes: Unit of observation is the driver. For every coaching date, the mean baseline performance and non-performance related variables of drivers who receive their first coaching (C) is compared to that of non-yet-coached colleagues (NC). Reported are the mean values over all coaching dates. Standard p -values as well as p -values that use a Bonferroni and Holm correction for multiple hypothesis testing are reported. Fuel economy in liters/100km; Performance on the ABC dimensions as the number of events/10km. Fewer events mean better driving behavior. Punctuality is the difference in minutes between actual and planned driving time. Number of passengers is based on check-ins with public transport cards. Morning and evening rush hours are from 7:00-10:00 and 16:00-19:00, respectively. Holiday rides take place during, for example, Christmas, New Year's Eve and school holidays. School rides are along routes with schools and universities as final destinations. Fill-ins are non-scheduled trips whereby a driver replaces a colleague from another base location.

***(*,*) : the corresponding p -values are less than 1% (5% or 10%).

economy and ABC events are estimated on the sample of trips completed by either a VDL or Intouro bus.

All 409 tenured drivers are included in our research design. Drivers with a temporary contract, 67 in total, are excluded because their behavior is only observed for short and irregular time spans. The trip-level observations in the final sample are matched with driver, trip, and daily weather characteristics.²⁶ We use this sample when we analyze the impact of coaching and peer-comparison feedback on a driver’s relative ranking. To keep the analysis succinct, we present full estimation results for fuel economy and acceleration in the main text and relegate some findings for braking and cornering events to the appendix. We will however highlight any important qualitative differences in treatment effects for acceleration and the outcomes braking and cornering when they arise.²⁷

2.5 Results

We first present the results of the written feedback program with the peer-comparison messages (2.5.1), followed by the effects of the in-person coaching program (2.5.2). Section 2.5.3 examines the interference between the two tailored feedback programs.

2.5.1 Feedback reports

This section reports the effect of the peer-comparison feedback program. To identify this effect, we estimate the following differences-in-differences (DID) regression specification:

$$\begin{aligned}
 Y_{its} &= \beta \cdot \text{postannounce}_i \\
 &+ \sum_{j=1}^4 I\{T_i = j\} \cdot (\tau_j \cdot \text{postfeedback}_{it} + \gamma_j \cdot \text{postexperiment}_{it}) \\
 &+ X_{its} \cdot \theta + \mu_i + \kappa_b + \nu_t + \zeta_{bt} + \xi_r + v_{its}.
 \end{aligned} \tag{2.1}$$

²⁶Appendix section 2.A.7 details the steps we have taken to construct the final sample.

²⁷From the ABC dimensions, we selected acceleration because of the higher average number of events in this dimension (Table 2.2) and the absence of intermediate changes in threshold settings (Table 2.A.6).

The dependent variable Y_{its} is the outcome variable of interest (fuel economy or ABC), indexed by driver (i), time in days (t), and the bus trip (s). In addition, the specification includes a vector X_{its} that contains the control variables listed in Table 2.1. A rich set of dummy variables controlling for driver (μ_i), bus type (κ_b), day (v_t), bus type interacted with day (ζ_{bt}), and route (ξ_r) fixed effects completes the specification.²⁸ Throughout, we use robust standard errors clustered at the driver level to account for within-driver correlation patterns in the error term (Bertrand, Duflo, and Mullainathan 2004). Importantly, because coaching takes place in parallel to feedback, a post-coaching dummy variable is included in the controls.

The dummy variable $postfeedback_{it}$ takes on the value one when the first feedback report has been delivered to driver i and is zero otherwise. This definition makes no selection on the actual reading of the report. From a policy perspective, this is useful because it captures the aggregate performance of the treatments when applied to an eligible population (Allcott 2011).²⁹ The dummy variable $postannounce_{it}$ equals one once a driver is informed about the upcoming EcoManager campaign, and zero otherwise; the dummy $postexperiment_{it}$ equals one once the report with the final notification message has been received, and zero otherwise.

The treatment indicator $T_i = j$ indicates when driver i is assigned treatment j , $j = 1, \dots, 4$. The τ -coefficients then estimate the treatment-specific effects of receiving tailor-made peer-comparison feedback, while the γ -coefficients measure the impact the withdrawal of peer-comparison messages has on performance (Barankay 2012). The β -coefficient captures the aggregate effect of the launch of the campaign and the switching on of the LED-arrays (which happen at the same date) on driving behavior.

Table 2.6 presents the results. Our preferred specification is reported in columns (2)-(4) and (6)-(8) and controls for being coached and time-variant

²⁸By interacting day- and bus type fixed effects, we relax the common trends assumption between bus types to address potential differences over time in the ease (or difficulty) of avoiding ABC events due to different thresholds per bus type. Of course, regressions with day fixed effects do not include the post announcement and post feedback dummies.

²⁹The start of the post-feedback period may differ per driver due to absence in the month on which the first report is based. A no-report indicator captures drivers operating after 15 December 2015 (first feedback round) but who have not yet received their first report.

driving features, such as weather conditions and the number of passengers. For fuel economy, we find a strong and significant reduction of $\beta=0.41$ liters/100km ($0.40\sigma_r$, $p < 0.001$) following the start of EcoManager. This is the joint effect of the launch-event and the switching on of the LED-arrays in the buses. The distribution of feedback reports generates an additional reduction of, on average, 0.13 liters/100km ($0.13\sigma_r$, $p = 0.105$). Column (3) shows that these estimates remain qualitatively unchanged but become more significant once driver fixed effects are added.

How does the experimental variation in the dosage of the number and nature of peer-comparison feedback messages affect worker productivity? Reassuringly, Table 2.6 shows no differences in fuel economy between the different treatment groups *before* the first feedback report is distributed. The estimates for the post-feedback dummy variable interacted with treatment indicators show no significant effect of the peer-comparison messages in the text boxes in addition to the general effect generated by the feedback reports. The point estimates across treatments for fuel economy are small in size, ranging from -0.11 to 0.05 liters/100km and are individually and jointly insignificant.

While fuel economy is an important outcome variable, the experimental feedback messages do not mention fuel economy but dissect a driver's relative performance into his/her performance on the disaggregate comfort dimensions acceleration, braking and cornering. The absence of a treatment effect for fuel economy need not imply that there is no effect at these 'lower' levels of driving behavior that are explicitly targeted by the intervention. Table 2.6 reveals that the pattern of effects for acceleration resembles the pattern for fuel economy: a large and significant effect following the announcement ($0.52\sigma_r$), a significant but smaller effect when the feedback reports are received ($0.35\sigma_r$), but again no indication that the text-box variation in the number and nature of peer-comparison messages matters. For braking and cornering, the estimates of the announcement and reception of feedback are $1.23\sigma_r$, $0.00\sigma_r$, and $0.14\sigma_r$, $0.10\sigma_r$, respectively (Table 2.A.8).³⁰ Treatment effects are also not seen on these dimensions.

³⁰The larger effect on braking is partly due to a change in the threshold setting for braking-events happening around the same time, see Appendix section 2.A.9.

Table 2.6: Targeted Peer-Comparison Feedback Effects on Driving Performance

Dependent variable:	Fuel Economy			Acceleration				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post-announcement	-0.305*** (0.062)	-0.411*** (0.053)	-0.367*** (0.038)		-2.250*** (0.129)	-1.154*** (0.116)	-1.229*** (0.094)	
T2 (1n/0p)	-0.258* (0.154)	-0.201 (0.125)			-0.328 (0.277)	-0.348 (0.249)		
T3 (1n/1p)	-0.040 (0.152)	-0.079 (0.172)			0.160 (0.362)	0.025 (0.305)		
T4 (3n/0p)	-0.081 (0.164)	-0.121 (0.134)			0.243 (0.284)	0.167 (0.256)		
Post-feedback	-0.402*** (0.077)	-0.130 (0.080)	-0.173*** (0.062)		-0.656*** (0.159)	-0.824*** (0.168)	-0.693*** (0.142)	
Post-feedback × T2 (1n/0p)	-0.105 (0.116)	-0.090 (0.092)	-0.026 (0.083)	-0.026 (0.082)	0.045 (0.238)	0.100 (0.216)	-0.097 (0.218)	-0.110 (0.215)
Post-feedback × T3 (1n/1p)	-0.013 (0.096)	0.047 (0.079)	0.036 (0.080)	0.023 (0.078)	-0.061 (0.227)	0.083 (0.200)	0.027 (0.199)	0.039 (0.196)
Post-feedback × T4 (3n/0p)	-0.082 (0.103)	0.017 (0.075)	-0.021 (0.074)	-0.032 (0.072)	-0.237 (0.231)	-0.117 (0.204)	-0.252 (0.208)	-0.236 (0.207)
Post-experiment	0.574*** (0.089)	0.554*** (0.062)	0.553*** (0.055)		0.367*** (0.142)	0.142 (0.119)	0.187* (0.104)	
Post-experiment × T2 (1n/0p)	-0.010 (0.130)	0.018 (0.093)	0.033 (0.079)	0.036 (0.077)	0.114 (0.220)	0.016 (0.187)	0.032 (0.149)	0.029 (0.149)
Post-experiment × T3 (1n/1p)	0.048 (0.114)	0.070 (0.081)	0.027 (0.070)	0.040 (0.069)	0.081 (0.188)	0.151 (0.158)	0.065 (0.137)	0.055 (0.138)
Post-experiment × T4 (3n/0p)	-0.056 (0.115)	0.050 (0.084)	0.033 (0.079)	0.023 (0.076)	0.017 (0.198)	0.157 (0.162)	0.080 (0.140)	0.060 (0.140)
Constant	24.639*** (0.110)	22.056*** (0.104)	22.546*** (0.080)	23.984*** (0.181)	10.790*** (0.217)	8.594*** (0.194)	11.868*** (0.191)	11.863*** (0.259)
R ²	.0148	.417	.499	.515	.0527	.405	.558	.573
Number of trip-level observations	533172	533172	533172	533172	352887	352887	352887	352887
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Weather dummies	No	Yes	Yes	No	No	Yes	Yes	No
Driver fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Day fixed effects	No	No	No	Yes	No	No	No	Yes
Bus type × day fixed effects	No	No	No	Yes	No	No	No	Yes

Notes: Identification of the written feedback treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is one in the time period from November 9, 2015, onwards (Kiddo-Event), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 13 December 2015 and after. The post-experimental period starts at November 15, 2016, when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in litres/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, non-scheduled rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after December 15, 2016, but who have not yet received their first report.

***, **, * : statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

In sum, with the exception of braking, the launch of the feedback program and the distribution has a significantly positive impact on fuel economy and all ABC dimensions. Table 2.A.10 shows that results remain significant at the $p = 0.05$ -level when we apply corrections for multiple hypothesis testing.³¹

The absence of an effect of peer-comparison feedback on conservation efforts among workers is consistent with findings in Blader et al. (2020). They note however that a focus on aggregate effects may mask temporal effects and improved performance among subgroups of drivers. In our case, estimates of the effects of peer-comparison feedback for each month separately do not suggest the presence of such temporal effects.³² There is no indication that drivers respond differently to the first peer-comparison messages that they receive than to the ones received in later months, for example because they lose attention. What about the possibility that certain subgroups of drivers are more responsive to the peer-comparison feedback program than others? Given our design feature that only the subset of drivers who actually belong to the top-25% or bottom-50% receive a message, it is indeed possible that we overlook some treatment effects among the subgroups that are treated. The treatment estimates presented so far estimate the overall effect of being assigned a peer-comparison feedback treatment condition. This intention-to-treat (ITT) estimate is a conservative estimate of the average effect of actually receiving positive or negative messages. For example, every month only about 70% of all drivers in treatments T2[1n0p] and T4[3n0p] actually receive messages in their textbox.³³

In an exploratory analysis, we group drivers on the basis of their per-

³¹At first sight, the fuel economy estimates for ‘post-experiment’ seem to suggest that the withdrawal of peer-comparison messages in the text box completely reverses the improvement in fuel economy achieved by the introduction of the EcoManager program: $[0.554] \sim [0.411 + 0.130]$. However, caution is needed in drawing this conclusion because the post-experimental period is relatively short and the specifications lack day fixed effects to absorb the unobserved day-to-day fluctuations in driving conditions. Also, the post-experiment coefficients for the ABC outcomes do not reflect a rebound effect.

³²Appendix section 2.A.11 contains the coefficient plots of these estimates.

³³In treatment T3[1n1p], about all drivers (97%) have their text box filled with a message, but with variation in whether the box contains only negative feedback (54%), only positive feedback (25%) or a combination of positive and negative feedback (21%). See Table 2.A.4 for detailed information on the composition of messages by treatment.

formance in month m (being in the top-25% or bottom-50%) and (for each group and ABC outcome dimension separately) regress a driver's ranking in month $m + 1$ on a dummy variable on receiving relative performance feedback on month m performance. The coefficient estimates (reported in Table 2.A.11) for cornering are all insignificant. For acceleration and braking, the coefficients for drivers in the bottom-50% are negative and in some cases significant, suggesting that the feedback messages help them to improve their ranking; for drivers in the top-25%, the coefficients are consistently and significantly positive, indicating that their average ranking deteriorates when having received positive feedback.

2.5.2 In-person coaching

To identify the effect of a single on-the-road coaching session on productivity outcomes in the weeks following coaching, we estimate the following DID regression specification:

$$\begin{aligned}
 Y_{its} = & \delta_0 I\{t = t_i^c\} + \sum_{\tau=1}^{10} \delta_{\tau} I\{t - t_i^c \in (7(\tau - 1), 7\tau]\} + \delta_{10}^+ I\{t - t_i^c > 70\} \\
 & + \sum_{\tau=1}^{10} \gamma_{-\tau} I\{t - t_i^c \in [-7\tau, -7(\tau - 1))\} \\
 & + X_{its} \cdot \theta + \mu_i + \kappa_b + v_t + \zeta_{bt} + \xi_r + \epsilon_{its}.
 \end{aligned} \tag{2.2}$$

As before, the dependent variable Y_{its} denotes the outcome of interest, the same set of control variables as in equation (2.1) is included and standard errors are clustered at the driver level. Day t_i^c denotes the specific day at which driver i is coached; recall that because of the phase-in design of the coaching program, drivers are coached at different days. The regressors include indicator functions $I(\cdot)$ to estimate the impact of coaching at: *a*) the day of coaching (coefficient δ_0); *b*) the first ten weeks following coaching (post-coaching coefficients $\delta_1, \dots, \delta_{10}$), and *c*) the ten weeks preceding coaching (pre-coaching coefficients $\gamma_{-1}, \dots, \gamma_{-10}$). The coefficient δ_{10}^+ absorbs any impact of coaching more than 10 weeks after the day of

coaching.³⁴

Figure 2.2 shows the temporal effects of coaching by plotting the pre- and post-coaching coefficients effects for the fuel economy and ABC outcomes. For fuel economy and acceleration, we observe a strong and immediate effect of coaching: on the day of coaching the fuel need reduces by 0.6 liters/100km ($0.58\sigma_r$) and the number of acceleration events by 1.1 events/10km ($0.50\sigma_r$). These effects persist for about seven to nine weeks. This suggests that, as time progresses, coaching effects decay and drivers seem to fall back into old driving habits.³⁵ We also observe an effect of coaching for braking and cornering, but these effects are much less pronounced and not (braking) or less (cornering) significant because of the lower baseline number of events (see Table 2.2). For none of the outcomes we observe differences in driving behavior in the 10 weeks prior to coaching, which lends support to our earlier conclusion that the selection for a coaching session is quasi-random and not based on prior performance.

Table 2.7 reports the main coefficients of regression specifications that take the entire period preceding coaching as the baseline period. Next to the standard p -values, we also report p -values that apply a Bonferroni and a Holm correction for multiple hypothesis testing (MHT). These are conservative methods to adjust for the fact that we consider the impact of coaching on four different outcome variables and three different time periods.³⁶

The regressions reveal that the largest improvements are observed on the day of coaching and with all adjusted p -values < 0.02 : fuel economy

³⁴Observations more than 10 weeks before coaching are the omitted period, the estimated δ -coefficients in Figure 2.2 show the average effect relative to this baseline period.

³⁵This fits into the body of evidence showing that in many domains, it is hard to induce persistent changes in habits (Brandon, Ferraro, List, Metcalfe, Price, and Rundhammer 2017).

³⁶The Bonferroni multiplicity-adjusted p -values are obtained by multiplying the unadjusted p -values by the number of hypotheses (12); the Holm multiplicity-adjusted p -values are obtained by ranking the unadjusted p -values from largest to smallest and to multiply each unadjusted p -value with its rank. Due to our stratified design, we cannot apply the less conservative MHT correction method developed by List, Shaikh and Xu (2019) that assumes simple random matching. When the joint dependence between the individual test statistics is positive (which is likely in our case given the positive correlations in Table 2.2) the latter method has a greater ability to detect false null hypotheses than the Bonferroni and Holm method.

Figure 2.2: Temporal Effects of In-Person Coaching

Notes: Driving performance in the 10 weeks before and after coaching based on trips with VDL and Intouro buses. The day of coaching itself is point 0 on the x -axis. The vertical spikes indicate 95% confidence intervals. The dependent variable fuel economy is measured in liters/100km and acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, number of passengers and bus stops, dummies for non-scheduled rides, additional coaching sessions, bus types, morning and evening rush hours, and the interaction of bus type and day fixed effects. Coaches themselves are excluded from the sample.

improves by 0.61 liters/100km ($0.55\sigma_r$), acceleration, braking and cornering by $0.48\sigma_r$, $0.11\sigma_r$ and $0.10\sigma_r$, respectively. For all outcomes except braking, we identify a short-run persistence effect in the first week following coaching. Only for acceleration, an effect is identified for the entire post-coaching period.

Table 2.7: In-Person Coaching Effects on Driving Performance

Outcome		Coefficient	p -values		
			Unadj.	Bonf.	Holm
Fuel economy	day of coaching	-0.6086	0.0000***	0.0000***	0.0000***
	day 1-7 after coaching	-0.3124	0.0000***	0.0000***	0.0000***
	post-coaching	-0.1235	0.0311**	0.3732	0.2799
Acceleration	day of coaching	-1.0803	0.000***	0.0000***	0.0000***
	day 1-7 after coaching	-0.6885	0.0000***	0.0000***	0.0000***
	post-coaching	-0.4469	0.0002***	0.0024***	0.0012***
Braking	day of coaching	-0.1094	0.0013***	0.0156**	0.0091***
	day 1-7 after coaching	-0.0244	0.3264	1.0000	1.0000
	post-coaching	-0.0159	0.5294	1.0000	1.0000
Cornering	day of coaching	-0.1901	0.0000***	0.0000***	0.0000***
	day 1-7 after coaching	-0.1043	0.0027***	0.0324**	0.0216**
	post-coaching	-0.0473	0.2118	1.0000	1.0000

Notes: Identification of in-person coaching effects on driving performance. The time period under consideration is the period for which we have complete logs available from all coaches (01/01/2015-30/04/2016). The dependent variable fuel economy is measured in liters/100km; acceleration, braking and cornering as the number of events per 10 kilometers. Reported are the DID effects of coaching on the day of coaching, in the first week following coaching and in the entire post-coaching period following coaching. Standard p -values as well as p -values that use a Bonferroni and Holm correction for multiple hypothesis testing are reported. Standard errors are clustered by driver. Full regression results are reported in Tables 2.A.12 and Table 2.A.13 in the appendix.

***(*, *): the corresponding p -values are less than 1% (5% or 10%).

Robustness check: heterogeneity in coach quality Coaching is provided by a small number of six coaches. One potential worry then is that the observed average treatment effects are not caused by an inherent feature of in-person coaching that is independent of who coaches, but is instead due to one or two coaches with idiosyncratic coaching qualities that are hard to copy. In that case, the data would not allow us to draw the general conclusion that in-person coaching improves worker productivity. We cannot directly compare differences in the way our six coaches provided feedback to drivers because we lack this information. We can however estimate the

treatment effect of each individual coach by considering the subsample of drivers instructed by that coach. When there is substantial heterogeneity in the quality of instructions given by the coaches, this should result in between-coach differences in treatment effects.

Figure 2.3: Temporal Effects of In-Person Coaching at Coach Level: Fuel Economy

Figure 2.3 shows these coach-level treatment effects of in-person coaching for the outcome fuel economy.³⁷ Despite the fact that the estimates are less precise due to the smaller subsamples, the pattern is remarkably consistent across coaches: on the day of coaching, for all coaches the point estimate of fuel savings is in the range $[-0.7, -0.4]$ liters/100km.³⁸ The diminishing and eventually vanishing of this effect in the seven to nine weeks following coaching is common to all coaches. Based on this evidence, we conclude that the observed effect can be attributed to features inherent to in-person coaching.

³⁷Figures 2.A.8-2.A.10 in the appendix show the coach-level treatment effects for the ABC dimensions.

³⁸Estimates for coach # 3 are ignored. These estimates are very imprecise because this coach operates in the urban area with IRIS buses for which fuel economy is not recorded.

Treatment heterogeneity Next we address whether there is heterogeneity in driver responses to coaching. The literature on peer effects in educational outcomes suggests that the effect of coaching may be heterogeneous, depending on a driver's own past performance. In this section, we address the open question whether this result carries over to non-educational contexts. We take the following non-parametric approach. For a driver coached in month m , we compare the driver's relative performance in productivity outcome y the month before ($m - 1$) and the month ($m + 1$) following coaching. We thus ignore a driver's relative performance in the month in which (s)he has been coached. We do this for all four productivity measures. Of course, because of reversion to the mean, there is a tendency for drivers who by chance attain a particularly high (low) ranking in month $m - 1$ to have a lower (higher) ranking in month $m + 1$. To account for this statistical phenomenon, we use the change in ranking non-coached drivers experience from month $m - 1$ to $m + 1$ as a benchmark against which we evaluate the change in ranking of drivers coached in month m .

Figure 2.4 plots for both groups the change in ranking. For non-coached drivers, the shaded area represents the local polynomial estimates of the relation between the ranking in months $m - 1$ and $m + 1$, along with a 95% confidence interval.³⁹ We fit separate polynomials for non-coached drivers part of the top-25%/bottom-50%/remaining group in month $m - 1$. Due to the reversion to the mean effect, the slope of each of these polynomials is less than one. The plots show clear evidence of heterogeneity in the effects of coaching: only drivers at the bottom half of the performance distribution benefit from coaching.⁴⁰ This result holds independent of which productivity outcome is considered (fuel economy or either of the ABC dimensions). The direction of our result is in contrast to the empirical literature on peer effects in education, which predominantly finds that high-achieving students benefit most from the presence of high-achieving peers. Possible explanations are that high-achieving workers in our setting have little room left for further improvement or are less open to a colleague's feedback.

³⁹In calculating the weighted local estimate, we use the standard Epanechnikov kernel function.

⁴⁰See Table 2.A.14 for regression estimates.

Figure 2.4: Differential Treatment Effects of Coaching

Notes: Green (blue): Average ranking of drivers (not) coached in month m in month $m - 1$ (x -axis) and month $m + 1$ (y -axis). The lines plot the non-parametric piece-wise local polynomial fit using an Epanechnikov kernel function. For non-coached drivers, the 95% confidence interval (shaded area) is shown as well.

2.5.3 Treatment interaction

We conclude with an exploratory analysis on the possible complementarity between in-person coaching and peer-comparison feedback. For this, we utilize the fact that a subset of drivers received coaching before receiving written feedback, while others received one or more written feedback reports before being coached.

We first consider whether having received the general feedback reports makes drivers more or less responsive to coaching. For fuel economy and acceleration, Figure 2.5 compares the response to coaching by drivers who did not yet receive feedback on paper with those who did. Although the confidence intervals have become wider because of the smaller samples, a comparison of panels (a) with (b) and (c) with (d) shows for both groups a similar pattern in the effect of coaching, both on the day of coaching as well as in the subsequent weeks. Hence, the effect of in-person coaching does not depend on having received prior feedback on one's performance in written form. We also checked whether the impact of coaching is affected by the treatment variation in the number and nature of peer-comparison messages in the text box. In line with the non-significant effects of the variation in text-box messages discussed in Section 2.5.1, we find no effect (see Appendix section 2.A.16).

What about the opposite case: do drivers who did already receive in-person coaching respond differently to the peer-comparison feedback messages than those who did not yet receive coaching? To answer this question, we compare the response to the feedback reports by drivers who were coached before the arrival date of the first report (December 15, 2015) with the response by drivers who did not receive coaching at all. For both subsamples, we run the same regression specification as in the previous section. Table 2.8 shows the results. Coached drivers seem more responsive to the general feedback report ('post-feedback'). Of most interest is the difference in response to the peer-comparison text-box messages. The treatment variation in the number of peer-comparison messages does not generate any observable change in productivity among the group of coached drivers, similar to what we found earlier for the entire sample. However, in the group

Figure 2.5: Impact In-Person Coaching With and Without Having First Received Written Feedback Reports

of drivers that has not yet been exposed to coaching, varying the nature and intensity of feedback does seem to have an effect. Drivers in the treatment group exposed to the highest number of negative feedback messages [3n0p] improve significantly in acceleration ($p=0.003$), braking ($p=0.002$) and fuel economy ($p=0.041$) compared to non-coached drivers that do not receive any peer-comparison messages. For fuel economy and braking, we also find positive effects in the group that is exposed to up to one negative message [1n0p] but at lower levels of significance ($p=0.016$ and $p=0.077$, respectively). For none of the outcome variables we find a treatment effect for treatment [1n1p] that mixes negative and positive feedback.

Table 2.8: Impact of Peer-Comparison Feedback, Conditional on Coaching

Dependent variable:	Fuel Economy		Acceleration		Braking		Cornering	
	Coached (1)	Non-coached (2)	Coached (3)	Non-coached (4)	Coached (5)	Non-coached (6)	Coached (7)	Non-coached (8)
Post-announcement	-0.455*** (0.072)	-0.338*** (0.068)	-1.261*** (0.168)	-1.090*** (0.175)	-1.336*** (0.048)	-1.302*** (0.050)	-0.225*** (0.051)	-0.328*** (0.059)
Post-feedback	0.171* (0.100)	-0.023 (0.105)	-0.673*** (0.249)	-0.335 (0.202)	-0.159*** (0.051)	0.199*** (0.057)	-0.157** (0.066)	-0.098 (0.107)
Post-feedback×T2 [1n/0p]	0.117 (0.135)	-0.379** (0.155)	0.260 (0.356)	-0.409 (0.318)	0.101 (0.071)	-0.130* (0.073)	0.061 (0.111)	-0.047 (0.116)
Post-feedback×T3 [1n/1p]	0.082 (0.107)	-0.117 (0.181)	0.183 (0.284)	-0.512 (0.417)	0.056 (0.068)	-0.088 (0.106)	0.075 (0.090)	0.081 (0.119)
Post-feedback×T4 [3n/0p]	0.064 (0.101)	-0.304** (0.147)	0.242 (0.246)	-1.012*** (0.326)	0.095 (0.076)	-0.282*** (0.090)	0.139* (0.075)	-0.226 (0.137)
Post-experiment	0.537*** (0.082)	0.476*** (0.086)	0.220 (0.163)	0.161 (0.188)	0.066 (0.045)	-0.003 (0.042)	-0.026 (0.035)	-0.092* (0.051)
R ²	0.498	0.502	0.561	0.519	0.233	0.219	0.379	0.407
# trip-level observations	232597	136482	164593	94667	164593	94667	167632	95938

Notes: The estimated regression specification includes the same controls, weather dummies and driver fixed effects as the specification estimated in columns (3) and (7) of Table 2.6. Full regression estimates reported in Tables 2.A.15-2.A.18.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

In sum, in-person coaching and peer-comparison feedback seem to interfere in an asymmetric manner: coaching is effective independent of prior exposure to peer-comparison messages but prior coaching renders peer-comparison messages non-effective. One possible explanation is that in-person coaching trumps peer-comparison feedback: once drivers have met a coach who gave them detailed feedback on what they do right and wrong on a trip, they become insensitive to subsequent messages about their relative performance. Our evidence shows no need to limit negative feedback or to mix it with positive feedback.

2.6 Discussion and conclusion

Given our empirical estimates on the impact of targeted peer-comparison feedback and in-person coaching on worker productivity, we next discuss the possible channels through which these programs change drivers' behavior. Cassar and Meier (2018) present a theoretical framework in which they distinguish four factors that affect work meaning: the need for autonomy, competence and relatedness and the mission of an organization.⁴¹ Different features of the feedback programs may impact these four dimensions of work meaning. The announcement of EcoManager and the provision of feedback may help to align drivers' beliefs with the (social) mission of the firm. Corrective peer-comparison feedback can help to develop competence, but may also make a driver feel less competent. To avoid the latter, it may work to combine corrective feedback with positive feedback. Intensifying feedback may on the other hand also induce feelings of being monitored and a loss in autonomy. Finally, being coached by an experienced colleague may strengthen the social relation with colleagues, thus benefiting relatedness.

The announcement of EcoManager has a strong and positive impact on all four productivity measures. One possible channel is that the campaign makes the social mission of the firm salient, thereby increasing the workers' sense of job meaning. Similar effects have been recorded in fundraising contexts by Grant (2008). From a principal-agent perspective, another possibility is that the announcement triggers reputational concerns – despite the firm's assurance that the feedback information will not be used in formal evaluations (List 2003). In line with evidence that checklists improve worker productivity by serving as a “memory aid” (Jackson and Schneider 2015), it is also possible that the switching on of the LED-arrays in the bus, happening around the feedback announcement date, serves as a permanent memory aid to drive carefully.

Additional regressions show that the announcement effect on fuel economy and acceleration for drivers less than fifty years old is about twice

⁴¹The first three are psychological needs that have been identified by self-determination theory as essential for human motivation, see Ryan and Deci (2000) and references therein.

the size of that of other drivers and highly significant.⁴² Figure 2.6 illustrates this for fuel economy. The lines show per age, years-of-service or gender category, respectively, the week-dummy coefficients over time (the week before the announcement, week 40, is taken as the baseline). Clearly, younger drivers show a sharper response to the EcoManager launch and the gap with the older drivers never closes (panel (a)). Does this reflect cohort differences in learning or reputational concerns? In case of the latter, we might expect a similar gap to occur if we categorize drivers by the number of years that they are already at the company. We do not observe this (panel (b)), suggesting that the increased saliency of the company's objectives especially resonates with younger drivers.

We find a positive effect of receiving written feedback. The feedback may indeed help drivers to become more competent drivers. Of course, drivers may do better on the dimensions measured because they know that these are monitored by the company. There is the possibility that their performance on unmeasured dimensions of job performance such as friendliness will deteriorate. We have no information on this but also did not hear from the company that the number of complaints increased. No treatment effects for the peer-comparisons in the text-messages are identified, except that not-yet-coached drivers show a larger improvement if they receive the full amount of corrective feedback. When workers receive negative feedback on certain dimensions, it does not increase their performance on these dimensions if the corrective feedback is accompanied by positive feedback on other dimensions.

Different explanations are possible for the strong and immediate impact of coaching. Coaching may improve the driver's competences, improve the worker's alignment with the company's mission, and/or deepen the relatedness to colleagues. We cannot distinguish between these three. However, the decay path tells us that neither the improved alignment with firm objectives, nor the improvement to human capital, nor the closer connection to colleagues is permanent. A social pressure explanation does

⁴²The outcome variables are regressed on post-announcement, post-feedback and post-experiment dummies, each interacted with four about equally sized age-categories (aged < 50, 50–54, 55–59 or ≥ 60). Full regression estimates reported in Tables 2.A.19–2.A.21.

not fit the pattern because the decay is not immediate. It may be that coaching serves as a memory aid for drivers with limited attention. Hanna, Mullainathan and Schwartzstein (2014) find a similar result for farmers. Farmers change behavior “when presented with summaries that highlight previously unattended-to relationships in the data”. Their paper cannot tell whether this effect remains as their final follow-up is two months after farmers receive this information. Our evidence clearly points out that, at least for bus drivers, the effect is short-lived.

Comparing the written feedback and in-person feedback, it seems that in-person feedback is more powerful but that the repetitive character of the monthly written report is necessary to let the impact last. This is in line with Dusch, Evans, Eze-Ajoku and Macis (2017) and also resonates the recommendations of Oreopoulos and Petronijevic (2018) for feedback provision in educational settings.

Our analyses show how carefully designed non-pecuniary strategies can improve conservation efforts among workers.⁴³ One robust outcome of the quasi-experimental variation in being coached is that in-person coaching immediately improves a driver’s performance. Especially drivers with the lowest prior performance receive a boost. Although this effect is transient, the decay is not immediate. In the weeks following coaching, 17.52 liter of fuel is saved per coached driver, which amounts to €19.27, or €60 per day of coaching. This is less than the cost of freeing up an experienced driver to coach, but the benefits may outweigh the cost if the company or passengers sufficiently value the improved comfort or environmental gains. Written performance feedback comes at a lower marginal cost. The reduction in fuel consumption following the announcement and distribution of general feedback reports is 5/6th the effect of coaching on the day of coaching. The additional peer-comparison messages generally do not have any impact, and we find no evidence that managers need to restrain themselves from giving negative feedback.

⁴³Most existing studies on energy conservation focus on households. Findings from these studies may not extend to corporate settings where workers are in a principal-agent relationship and have no direct financial stake in energy conservation. Our findings contribute to a growing body of literature on non-financial interventions aimed at energy conservation in the workplace.

A few directions for future research follow from our findings. First, the increased adoption of digital monitoring technologies at the work floor creates many new opportunities for tailoring feedback to motivate workers. This chapter has examined the impact of different feedback channels and variation in feedback intensity, but we reckon that designing and evaluating other data-driven incentives could yield fruitful research. Second, we document important interaction effects between targeted peer-comparison feedback and in-person coaching, and believe that more research should be done to investigate interactions between non-financial incentives. Finally, and more in general, the question how conservation efforts can be stimulated when someone else pays the bill is in need of more answers.

Figure 2.6: Response in Fuel Economy to Written Feedback per Subgroup of Drivers

Notes: Coefficients week dummies on trips with VDL buses. The time period is from 01/01/2015 to 31/01/2017. The dashed lines indicate the launch date of the EcoManager program [09/11/2015], the distribution of the first feedback report [15/12/2015] and the distribution of the report with the final notification message [15/11/2016]. The dotted lines indicate the distribution of the intermediate monthly feedback reports. The first dotted line indicates the moment EcoManager promotion materials are sent to the locations [05/10/2015]. Controls include: weather conditions, travel distance, route dummies, number of passengers and bus stops, having received coaching, and dummies for bus type, morning and evening rush hours and fill-in rides. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

2.A Appendix

2.A.1 Estimates of bus driver fixed effects

Figure 2.A.1: Estimates of Driver Fixed Effects for Fuel Economy

Notes: Estimates based on trips completed with VDL and Intouro buses.

2.A.2 Model of the estimation of the residual standard deviation

Let the outcome variable of interest, Y_{it} (fuel economy or ABC), indexed by driver (i) and trip (t), be given by:

$$Y_{it} = X_{it} \cdot \beta + \mu_i + f(s_{it}) + g(z_{it}) + \epsilon_{it}. \quad (2.3)$$

In this specification, X_{it} includes all observable determinants of driver performance, day, driver and route fixed effect. The μ_i parameters denote driver fixed effects that absorb all time-invariant unobservables at the driver level and the functions $f()$ and $g()$ reflect the potential impact of having been exposed to the peer-comparison treatment (s_{it}) and coaching (z_{it}), respectively. Especially for the ABC-dimensions our baseline period before treatment is too small for a reliable estimation of the residual variation. Hence, similar to Chetty, Friedman and Rockoff (2014), we use the entire sample to compute driver's residual outcomes. The residual r_{it} in outcomes after accounting for the observable determinants of driver performance, day, driver and route fixed effect is constructed as follows:

$$r_{it} = Y_{it} - X_{it} \cdot \beta = \mu_i + f(s_{it}) + g(z_{it}) + \epsilon_{it}. \quad (2.4)$$

The standard deviation of the residual variation is computed as:

$$\sigma_r = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (r_i - \bar{r})^2},$$

with $r_i = \frac{1}{n_i} \sum_t^{n_i} r_{it}$ driver i 's mean residual over n_t trips and $\bar{r} = \frac{1}{n} \sum_t^n r_i$ the mean residual across drivers. In case the impact of treatment (written feedback or in-person coaching) is correlated with unobservables at the driver level, σ_r will over- or understate the residual variation across drivers. When for example coaches tend to select on average worse drivers (for which we do not find evidence) or when worse drivers benefit more from coaching (for which we do find some evidence), σ_r will be a conservative, downward biased estimate of the actual across driver variation in μ_i 's.

2.A.3 Determinants of ABC outcomes

Table 2.A.1: Determinants of Acceleration

Dependent variable:	Acceleration				
	(1)	(2)	(3)	(4)	(5)
VDL 10m	0.953*** (0.073)	0.969*** (0.073)	1.090*** (0.071)	1.139*** (0.072)	1.086*** (0.048)
VDL 14m	-0.080 (0.149)	-0.127 (0.146)	-0.189 (0.157)	-0.301 (0.191)	-0.172 (0.198)
Intouro	-1.454*** (0.210)	-1.414*** (0.213)	-1.756*** (0.205)	-1.472*** (0.214)	-1.235*** (0.224)
Rush hour 7-10am			-0.896*** (0.144)	-0.909*** (0.146)	-0.844*** (0.127)
Rush hour 4-7pm			1.182*** (0.157)	1.199*** (0.156)	1.146*** (0.111)
Non-scheduled trip			0.353* (0.202)	0.461** (0.208)	0.624*** (0.165)
No. of stops per km.			2.274*** (0.194)	2.376*** (0.188)	2.361*** (0.169)
Trip length (in km.)			-0.025*** (0.002)	-0.021*** (0.002)	-0.021*** (0.002)
Ln(No. of passengers)			1.858*** (0.081)	1.728*** (0.080)	1.648*** (0.072)
Punctuality			-0.039*** (0.015)	-0.031** (0.015)	0.049*** (0.007)
Constant	13.267*** (0.188)	14.002*** (0.199)	6.341*** (0.457)	4.694*** (0.429)	4.718*** (0.357)
R ²	.514	.516	.537	.543	.602
Number of trip-level observations	513866	513866	513866	513866	513866
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of acceleration events per 10 km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Standard errors are clustered by driver.

*** (**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.2: Determinants of Braking

Dependent variable:	Braking				
	(1)	(2)	(3)	(4)	(5)
VDL 10m	0.959*** (0.032)	0.969*** (0.032)	0.984*** (0.032)	1.037*** (0.032)	1.017*** (0.031)
VDL 14m	0.232*** (0.029)	0.195*** (0.033)	0.178*** (0.037)	0.145*** (0.055)	0.197*** (0.071)
Intouro	3.758*** (0.240)	3.776*** (0.240)	3.632*** (0.239)	3.741*** (0.236)	3.771*** (0.227)
Rush hour 7-10am			0.005 (0.054)	-0.008 (0.054)	-0.024 (0.033)
Rush hour 4-7pm			0.207*** (0.077)	0.221*** (0.076)	0.252*** (0.046)
Non-scheduled trip			-0.061 (0.080)	-0.053 (0.077)	-0.078 (0.085)
No. of stops per km.			0.275** (0.118)	0.324*** (0.113)	0.321*** (0.089)
Trip length (in km.)			-0.005*** (0.001)	-0.002*** (0.001)	-0.002*** (0.000)
Ln(No. of passengers)			0.384*** (0.037)	0.287*** (0.034)	0.280*** (0.031)
Punctuality			-0.030*** (0.006)	-0.025*** (0.006)	-0.000 (0.002)
Constant	0.956*** (0.066)	1.303*** (0.076)	-0.001 (0.253)	0.121 (0.220)	0.115 (0.185)
R ²	.463	.467	.473	.492	.575
Number of trip-level observations	513866	513866	513866	513866	513866
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of braking events per 10 km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Standard errors are clustered by driver.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.3: Determinants of Cornering

Dependent variable:	Cornering				
	(1)	(2)	(3)	(4)	(5)
VDL 10m	-0.085*** (0.018)	-0.079*** (0.018)	-0.110*** (0.018)	-0.078*** (0.019)	-0.064*** (0.013)
VDL 14m	-0.164*** (0.039)	-0.132*** (0.043)	-0.109** (0.051)	-0.125 (0.077)	-0.103 (0.117)
Intouro	-0.343*** (0.043)	-0.320*** (0.044)	-0.463*** (0.059)	-0.371*** (0.065)	-0.279*** (0.055)
Rush hour 7-10am			-0.130 (0.101)	-0.139 (0.099)	-0.172*** (0.049)
Rush hour 4-7pm			0.164 (0.116)	0.171 (0.115)	0.177*** (0.045)
Non-scheduled trip			-0.116 (0.119)	-0.108 (0.117)	-0.327** (0.159)
No. of stops per km.			-0.509*** (0.134)	-0.570*** (0.134)	-0.518*** (0.095)
Trip length (in km.)			-0.004*** (0.001)	-0.002** (0.001)	-0.003*** (0.000)
Ln(No. of passengers)			0.151*** (0.031)	0.098*** (0.032)	0.094*** (0.023)
Punctuality			-0.121*** (0.014)	-0.119*** (0.014)	-0.055*** (0.006)
Constant	3.096*** (0.163)	3.386*** (0.179)	3.797*** (0.318)	1.610*** (0.244)	1.467*** (0.177)
R ²	.5	.502	.506	.513	.676
Number of trip-level observations	513866	513866	513866	513866	513866
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of cornering events per 10 km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Standard errors are clustered by driver.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

2.A.4 Timing of coaching sessions

Figure 2.A.2: Time of First Coaching

Notes: Moment of first coaching for drivers. Dark blue bars indicate the drivers who received their first coaching during a specific week as a share of the total number of drivers operating during that week. The light blue bars depict the cumulative share of coached drivers operating during a week. Feedback was announced on Nov. 9, 2015 and first distributed as a monthly report after Dec. 15, 2015. Peer-comparison messages were removed from the reports from Nov. 2016 onwards.

Two things in Figure 2.A.2 related to the coaching program deserve some further explanation. First, the cumulative share of coached drivers operating during a week is more or less flat after April 2016. We have complete coach logs for the period till 30 April, 2016. Some coaches indicated that they no longer provided or kept track of coaching after April 2016. In our evaluation of the coaching program, we therefore restrict attention to the period until 30 April 2016. Second, 30 drivers (10% of all coached drivers) received coaching prior to the feedback announcement.

2.A.5 Sample feedback report

Figure 2.A.3 reproduces a specimen of the feedback report drivers received once a month between December 2015 - November 2016.

Figure 2.A.3: Sample Feedback Report

Notes: Confidential information (related to the driver) has been removed.

2.A.6 Driver exposure to targeted feedback

The tailored nature of the messages is illustrated in Table 2.A.4. Panel A reports the percentage share of drivers receiving one of the possible message combinations in each treatment and feedback round (conditional on receiving a feedback report).⁴⁴ It highlights the flexible design of the treatments. Drivers in the control group, treatment 1 (0n/0p), receive no tailored feedback messages. Each treated driver is assigned an individualized message combination which points to behaviors that require attention. In treatment 2 (1n/0p), for example, about 70% of the drivers receive a negative message in a given feedback round, meaning that they perform poorly compared to peers on one of the three comfort dimensions. The remaining 30% performs well on all dimensions and is therefore not notified with a message. Panel B details the composition of the message combinations and shows that all ABC dimensions are well-represented.

How often a treated driver is in the top-25% or bottom-50% on a given driving dimension is shown in Figure 2.A.4. The figure plots the number of feedback rounds a driver is in the bottom or top part of the reference group divided by the total number of feedback rounds in which the driver received a feedback report. This gives an indication how often a driver is eligible for targeted messages. For many drivers it varies per round whether they were in the target groups. On each dimension, we observe that there are drivers who were always or never in the bottom (top) part. On acceleration, 19% (16%) of the treated drivers were never (always) in the bottom 50%. For the top 25%, the corresponding figures are 42% (9%). Outcomes are similar for braking and cornering.

⁴⁴No report is created when drivers were absent in the previous month (on which the report is based).

Table 2.A.4: Incidence and Composition of Targeted Peer-Comparison Messages

Feedback round	T2 (1n/0p)		T3 (1n/1p)		T4 (3n/0p)	
	0n/0p	1n/0p	0n/0p	1n/1p	0n/0p	1n/0p
Panel A: incidence of messages						
December 2015	31%	69%	3%	53%	24%	20%
January 2016	27%	73%	7%	49%	22%	22%
February 2016	31%	69%	2%	49%	21%	28%
March 2016	30%	70%	3%	52%	23%	22%
April 2016	30%	70%	1%	53%	27%	19%
May 2016	29%	71%	2%	55%	27%	16%
June 2016	30%	70%	5%	51%	23%	22%
July 2016	26%	74%	3%	54%	24%	19%
August 2016	27%	73%	4%	47%	23%	25%
September 2016	32%	68%	2%	56%	26%	27%
October 2016	30%	70%	3%	52%	22%	19%
Panel B: composition of messages (A%;B%;C%)						
December 2015	(31;46;23)	(40;23;38)	(33;39;28)	(80;33;87)	(30;35;35)	(65;59;76)
January 2016	(33;33;33)	(26;30;45)	(38;33;29)	(76;29;95)	(30;30;39)	(69;69;63)
February 2016	(20;41;39)	(34;43;23)	(55;15;30)	(67;59;74)	(44;20;36)	(52;81;67)
March 2016	(32;43;25)	(34;32;34)	(32;23;45)	(90;52;57)	(27;33;40)	(65;70;65)
April 2016	(33;38;30)	(27;37;35)	(42;23;35)	(56;72;72)	(35;24;41)	(63;74;63)
May 2016	(38;29;32)	(34;38;28)	(31;27;42)	(100;56;44)	(29;29;41)	(64;73;64)
June 2016	(40;35;25)	(31;37;33)	(45;27;27)	(67;62;71)	(28;28;44)	(68;71;61)
July 2016	(34;31;34)	(33;24;43)	(57;22;22)	(61;78;61)	(32;27;41)	(64;76;60)
August 2016	(38;30;33)	(29;36;36)	(32;32;36)	(88;58;54)	(43;30;26)	(61;61;78)
September 2016	(41;29;31)	(28;33;39)	(32;36;32)	(73;80;47)	(28;44;28)	(69;56;75)
October 2016	(47;25;27)	(39;27;35)	(38;24;38)	(50;86;64)	(50;25;25)	(55;68;77)

Notes: Panel A reports the percentage share of drivers receiving one of the possible message combinations in each treatment (conditional on receiving a feedback report). Drivers do not receive a report when they were absent in the previous month (on which the report is based). Panel B shows the composition of these combinations in terms of the ABC comfort dimensions. Negative (positive) messages are provided if a driver belongs to the bottom 50% (top 25%) of a reference group of peers on one of the comfort dimensions. The reference group consists of drivers who share the same base location and treatment status. Drivers in the control group, T1 (0n/0p), receive no peer-comparison messages. A printed version of the feedback report (with the messages integrated) is created around the 15th day of each feedback round and delivered via the team manager or pigeonhole.

Figure 2.A.4: Share of Feedback Rounds in Top 25% or Bottom 50% for Treated Drivers

Notes: The figures show for each ABC driving dimension the distribution of feedback round shares in which treated drivers were in the top 25% or bottom 50% of the peer reference group. The shares are calculated as the number of feedback rounds a driver was in the bottom or top part of the reference group divided by the total number of feedback rounds in which a feedback report was constructed for the driver. It indicates how often a driver was eligible for a targeted peer-comparison message on a given driving dimension (the received message combination depends on the treatment condition). The reference group consists of drivers who share the same base location and treatment status.

2.A.7 Construction final sample

The initial sample consists of 1,278,913 trip-level observations. Table 2.A.5 summarizes the steps taken to arrive at the final sample that is used in our analysis. The absence of a bus identifier has been the most important reason to discard observations (22.7% of the initial sample). Without such an identifier, it is not possible to link the trip to the outcome data of interest (fuel economy and ABC dimensions). Observations are also dropped when there are technical mismatches between the bus type and EOBR data (about 5%) or when there are between-trip inconsistencies at the driver level (3%). We also exclude extreme outcomes regarding punctuality or outcomes that suggest a temporary technical recording problem in the EOBR (< 0.5 percent). These comprise observations of fuel economy being less than 12.5 or more than 100 (1,259 obs; 0.10%), a difference of more than one hour between actual and planned driving time (156 obs; 0.01%), and outcomes that are more than five standard deviations above the means of the ABC dimensions (4,003 obs; 0.31%).

Table 2.A.5: Cleaning Steps for Sample Construction

	% share of full sample	
Full sample	1,278,913	
<i>Reason for dropping observation:</i>		
Duplicate observation (in terms of all variables)	(6,762)	0.50
No bus identifier	(290,737)	22.73
Bus type not eligible for EOBR	(34,870)	2.73
Error message from EOBR	(31,118)	2.46
Within-driver obs. with the same departure date/time	(37,575)	2.94
Very short rides (less than 1 kilometer)	(29,588)	2.31
Punctuality: more than 1 hour	(156)	0.01
<i>Unreasonable outcomes of dependent variables:</i>		
- Fuel economy: less than 12.5 or more than 100	(1,259)	0.10
- ABC dimensions: more than 5 SDs above the mean	(4,003)	0.31
	842,845	

Notes: Fuel economy is measured in liters/100km. The ABC comfort dimensions are the number of events per 10 kilometers (fewer events mean better driving behavior). Trip punctuality is the difference between actual and planned driving time.

The company uses different bus types. For the trips in the final sample, Table 2.A.6 shows which bus type is used. The table reveals that the vast majority of trips ($> 63\%$) is completed with a VDL bus. More specifically, over half of the trips are completed with the VDL 12 meter bus. The IRIS

bus is used in about 30% of all trips. Most of these IRIS buses run on natural gas, which implies that no fuel economy score is recorded for these trips. These gas buses are almost exclusively used in the province's capital because of the lower CO₂ emissions of natural gas compared to diesel. A small minority (<10%) of trips is completed with a third bus type, the Mercedes Intouro.

Table 2.A.6: Fleet Information

Bus type	length (m)	fuel type	No. trips	% share of full sample
VDL AMBASSADOR ALE 106	10.6	diesel	76,815	9.11
VDL CITEA LLE 120	12.0	diesel	452,375	53.67
VDL CITEA XLE 145	14.5	diesel	5,830	0.69
IRISBUS CITELIS 10,5 M	10.5	diesel	22,882	2.71
IRISBUS CITELIS 10,5 M CNG	10.5	natural gas	70,341	8.35
IRISBUS CITELIS 12 M	12.0	diesel	41,048	4.87
IRISBUS CITELIS 12 M CNG	12.0	natural gas	113,731	13.49
MERCEDES BENZ INTOURO	13.0	diesel	59,823	7.10
Analysis set			842,845	
VDL			535,020	63.5
IRISBUS			248,002	29.4
INTOURO			59,823	7.1

Table 2.A.7 shows per bus type the start of data recording. Data on fuel economy are available for a somewhat longer time period than the outcomes on the ABC-comfort dimensions.

Table 2.A.7: Start Date of Data Recording per Bus Type

	VDL	Intouro	IRIS
Fuel consumption	1 Jan. 2015	1 Jan. 2015	n.a.
Acceleration	1 Sep. 2015	9 Nov. 2015	1 Sep. 2015
Braking	1 Sep. 2015	9 Nov. 2015	1 Sep. 2015
Cornering	1 Sep. 2015	1 Sep. 2015	9 Nov. 2015
No. of trips	535,020	59,823	248,002

Notes: The number of trips for which the ABC events are recorded is lower than the total number of trips mentioned in the table, because recording of ABC events commenced only in Sept. 2015. This explains the difference between the number of observations mentioned in this table and those in the regression tables of Section 2.5.

2.A.8 Results on braking and cornering

Table 2.A.8: Targeted Peer-Comparison Feedback Effects on Driving Performance: Braking and Cornering

Dependent variable:	Braking				Cornering			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post-announcement	-1.267*** (0.027)	-1.268*** (0.027)	-1.292*** (0.027)		-0.334*** (0.029)	-0.272*** (0.033)	-0.317*** (0.028)	
T2 (1n/0p)	-0.081 (0.059)	-0.084 (0.058)			-0.075 (0.117)	-0.080 (0.116)		
T3 (1n/1p)	-0.083 (0.063)	-0.082 (0.063)			-0.152 (0.121)	-0.158 (0.117)		
T4 (3n/0p)	-0.029 (0.065)	-0.030 (0.063)			0.012 (0.121)	0.007 (0.120)		
Post-feedback	0.070** (0.033)	0.020 (0.035)	0.000 (0.036)		-0.145*** (0.052)	-0.160*** (0.058)	-0.193*** (0.051)	
Post-feedback $\times$ T2 (1n/0p)	0.063 (0.050)	0.071 (0.049)	0.052 (0.051)	0.047 (0.049)	0.044 (0.070)	0.048 (0.070)	0.063 (0.068)	0.059 (0.068)
Post-feedback $\times$ T3 (1n/1p)	0.003 (0.054)	0.015 (0.052)	0.009 (0.053)	0.015 (0.051)	0.114* (0.069)	0.118* (0.067)	0.110* (0.065)	0.108* (0.064)
Post-feedback $\times$ T4 (3n/0p)	-0.017 (0.053)	0.001 (0.052)	-0.020 (0.051)	-0.014 (0.050)	0.035 (0.068)	0.040 (0.068)	0.063 (0.063)	0.062 (0.062)
Post-experiment	0.033 (0.028)	-0.003 (0.027)	-0.011 (0.027)		-0.100*** (0.034)	-0.106*** (0.036)	-0.051** (0.025)	
Post-experiment $\times$ T2 (1n/0p)	0.005 (0.040)	-0.004 (0.038)	0.009 (0.036)	-0.001 (0.036)	0.067 (0.043)	0.064 (0.042)	0.004 (0.033)	0.005 (0.033)
Post-experiment $\times$ T3 (1n/1p)	0.099** (0.039)	0.107*** (0.035)	0.099*** (0.035)	0.089*** (0.034)	0.034 (0.042)	0.043 (0.040)	-0.026 (0.032)	-0.029 (0.031)
Post-experiment $\times$ T4 (3n/0p)	0.080** (0.039)	0.059 (0.039)	0.058 (0.039)	0.041 (0.038)	0.068 (0.052)	0.076 (0.050)	-0.007 (0.040)	-0.004 (0.040)
Constant	1.751*** (0.050)	1.367*** (0.055)	2.313*** (0.059)	2.157*** (0.085)	1.142*** (0.090)	1.302*** (0.096)	1.637*** (0.056)	1.735*** (0.078)
R ²	.0841	.2	.224	.28	.0264	.0912	.388	.397
Number of trip-level observations	352887	352887	352887	352887	359066	359066	359066	359066
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Weather dummies	No	Yes	Yes	No	No	Yes	Yes	No
Driver fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Day fixed effects	No	No	No	Yes	No	No	No	Yes
Bus type $\times$ day fixed effects	No	No	No	Yes	No	No	No	Yes

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015, until 31/01/2017. Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is equal to one in the period from 09/11/2015 onwards, zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variables braking and cornering are measured as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report. ***(*, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

2.A.9 Intermediate changes in ABC settings

During the period of data collection, the company changed some of the threshold settings for the ABC dimensions. An increase (decrease) in the threshold has the effect of reducing (increasing) the number of recorded events. The company has provided a detailed list of when which threshold has been changed on which bus type(s). This list is presented in Table 2.A.9.

We note that the VDL buses did not have any changes in settings throughout the period of data collection, except for the braking threshold, which was increased on Oct. 16, 2015 and subsequently slightly decreased on Nov. 5. Especially the Oct. 16 change seems to result in a drop of the number of recorded events, as can be seen in Figure 2.A.5b, which shows by bus type the development of the scores in ABC dimensions and fuel economy (weekly averages).

As Figure 2.A.5a-c (right axis) show, the IRIS buses record an importantly larger number of events in all three dimensions than the other two bus types. This is due to the fact that this is the bus type of choice in the province's capital, where the routes are characterized by many road bends and stops. Unfortunately, both the acceleration and cornering thresholds for the IRIS buses were increased at Dec. 11, just around the time the drivers received their first report. We therefore cannot identify whether the drop we observe in Figures 2.A.5b and c for the number of braking and cornering events is due to the report or the change in settings. For cornering, no data are available for the period prior to the kickoff event. Due to these issues, we exclude the IRIS bus from the analysis in the main text.

The Intouro buses experienced a number of recalibrations, but all before the date the first feedback report was received by the drivers and mostly comprised only two to four buses out of a total of 29 Intouro buses. Because calibration for these buses was late, we have no records for acceleration and braking for the period prior to the kickoff event, as Figures 2.A.5a and b clearly show.

Other than the ABC dimensions, the fuel economy records of both the VDL and Intouro buses cover the entire period from January 2015 till

January 2017 and have not been subject to any change in measurement. For completeness, Figures 2.A.5d show the weekly average fuel economy for both the VDL and Intouro buses.

Table 2.A.9: Change in Threshold Settings

Bus type	date of change	dimension affected	nature of threshold change
IRISBUS 12 M	Dec. 11, 2015	A	decrease
IRISBUS 12 M CNG	Dec. 11, 2015	A	decrease
IRISBUS 10,5 M	Dec. 11, 2015	C	increase
IRISBUS 10,5 M CNG	Dec. 11, 2015	C	increase
VDL AMBASSADOR ALE 106	Oct. 16, 2015	B	increase
	Nov. 5, 2015	B	decrease
VDL CITEA LLE 120	Oct. 16, 2015	B	increase
MERCEDES BENZ INTOURO			
bus no. 7503, 7504	Sept. 16, 2015	ABC	recalibration
bus no. 7503, 7504	Oct. 02, 2015	ABC	recalibration
bus no. 7501, 7502, 7503	Nov. 25, 2015	ABC	recalibration
all	Dec. 03, 2015	AB	recalibration

Figure 2.A.5: Development Over Time in ABC Dimensions (No. Events) and Fuel Economy (liters/100km) – Weekly Averages by Bus Type

Notes: Averages are calculated based on all trips in the analysis set. The red dashed lines indicate the launch date of the EcoManager program [09/11/2015], the distribution of the first feedback report [15/12/2015] and the distribution of the report with the final notification message [15/11/2016]. The dotted lines indicate the distribution of the intermediate monthly feedback reports. The first dotted line indicates the moment EcoManager promotion materials are send to the locations [05/10/2015].

2.A.10 Treatment effects targeted peer-comparison feedback, MHT correction

Table 2.A.10: Written Feedback on Driving Performance, Multiple Hypotheses Testing Correction

Outcome		Δ	p -values		
			Unadj.	Bonf.	Holm
Fuel Economy	Post-announcement	-0.3674	0.0000***	0.0000***	0.0000***
	Post-feedback	-0.1726	0.0053***	0.0635*	0.0423**
	Post-experiment	0.5529	0.0000***	0.0000***	0.0000***
Acceleration	Post-announcement	-1.2294	0.0000***	0.0000***	0.0000***
	Post-feedback	-0.6928	0.0000***	0.0000***	0.0000***
	Post-experiment	0.1866	0.0732	0.879	0.6592
Braking	Post-announcement	-1.2917	0.0000***	0.0000***	0.0000***
	Post-feedback	0.0005	0.9898	1.0000	1.0000
	Post-experiment	-0.0111	0.6855	1.0000	1.0000
Cornering	Post-announcement	-0.3168	0.0000***	0.0000***	0.0000***
	Post-feedback	-0.1931	0.0001***	0.0017***	0.001***
	Post-experiment	0.0629	0.3588	1.0000	1.0000

Notes: Identification of written feedback on driving performance. The dependent variable fuel economy is measured in liters/100km; acceleration, braking and cornering as the number of events per 10 kilometers. The time period under consideration is from 01/01/2015 until 31/01/2017. Standard p -values as well as p -values that use a Bonferroni and Holm correction for multiple hypothesis testing are reported. Standard errors are clustered by driver. Full regression results are reported in Tables 2.6 and Table 2.A.8.

***(**, *) : the corresponding p -values are less than 1% (5% or 10%).

Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is one in the period from November 9, 2015, onwards (kickoff-event), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15 December 2015 and after. The post-experimental period starts at November 15, 2016, when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by

driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, non-scheduled rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after December 15, 2015, but who have not yet received their first report.

2.A.11 Intertemporal treatment differences

Figure 2.A.6 examines temporal effects by plotting the treatment effects per feedback round. The first round starts around 15 December 2015, with a new report being distributed in each subsequent month. The feedback report in November 2016 contains a text message notifying all treated drivers that they will no longer receive peer-comparison messages. The general pattern is that there are no intertemporal effects of the peer-comparison messages on driving behavior.

Figure 2.A.6: Intertemporal Treatment Differences Targeted peer-comparison Feedback [Fuel Economy and Acceleration]

Notes: Treatment effects per feedback round based on trips with VDL and Intouro buses. The time period is from 01/09/2015-31/01/2017. **Pre:** 01/09/2015-08/11/2015 [pre-announcement period]; **A:** 09/11/2015-14/12/2015 [announcement period]; **F*i*:** 15/12/2015-14/11/2016 [feedback period]; **P*i*:** 15/11/2016-31/01/2017 [post-experiment period]. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours and fill-in rides. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

Figure 2.A.7: Intertemporal Treatment Differences Targeted Peer-Comparison Feedback [Braking and Cornering]

Notes: Treatment effects per feedback round based on trips with VDL and Intouro buses. The time period is from 01/09/2015-31/01/2017. **Pre:** 01/09/2015-08/11/2015 [pre-announcement period]; **A:** 09/11/2015-14/12/2015 [announcement period]; **F*i*:** 15/12/2015-14/11/2016 [feedback period]; **P*i*:** 15/11/2016-31/01/2017 [post-experiment period]. The dependent variables braking and cornering are measured as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours and fill-in rides. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

2.A.12 Change in ranking following relative performance feedback

Table 2.A.11: Change in ranking month $(m + 1)$ vs. m following relative performance feedback on performance indicator in month m

Dep. var		Rank month ($m + 1$)			
Acceleration					
	Message	0.001 (0.02)	-0.007 (0.016)	-0.086** (0.036)	0.054** (0.022)
	Rank month m	0.725*** (0.079)	0.644*** (0.053)	0.851*** (0.075)	1.105*** (0.22)
	obs.	464	521	487	197
	R^2	0.2503	0.2195	0.2884	0.214
Sample	Rank month m :	bottom-50%	bottom-50%	bottom-50%	top-25%
	Treatment:	T2[0p1n]	T3[1p1n]	T4[0p3n]	T3[1p1n]
Braking					
	Message	-0.043* (0.022)	-0.044** (0.019)	-0.063 (0.048)	0.113*** (0.036)
	Rank month m	0.704*** (0.096)	0.723*** (0.062)	0.712*** (0.113)	0.581** (0.253)
	obs.	468	516	492	204
	R^2	0.1928	0.216	0.1635	0.1248
Sample	Rank month m :	bottom-50%	bottom-50%	bottom-50%	top-25%
	Treatment:	T2[0p1n]	T3[1p1n]	T4[0p3n]	T3[1p1n]
Cornering					
	Message	-0.003 (0.012)	-0.008 (0.012)	-0.014 (0.031)	0.012 (0.019)
	Rank month m	0.911*** (0.043)	0.911*** (0.049)	0.808*** (0.062)	0.907*** (0.148)
	obs.	470	508	500	196
	R^2	0.5877	0.524	0.4244	0.2048
Sample	Rank month m :	bottom-50%	bottom-50%	bottom-50%	top-25%
	Treatment:	T2[0p1n]	T3[1p1n]	T4[0p3n]	T3[1p1n]

Notes: All regressions include a constant. Standard errors clustered by driver.

2.A.13 In-person coaching effects on driving performance

Table 2.A.12: In-Person Coaching Effects on Driving Performance: Fuel Economy and Acceleration

Dependent variable:	Fuel Economy			Acceleration	
	(1)	(2)	(3)	(4)	(5)
Post-coaching	-0.168*** (0.064)	-0.124** (0.057)		-0.468*** (0.134)	-0.447*** (0.120)
Day of first coaching			-0.609*** (0.085)		-1.080*** (0.165)
1–7 days after			-0.312*** (0.058)		-0.689*** (0.120)
8–14 days after			-0.174*** (0.064)		-0.467*** (0.137)
15–21 days after			-0.278*** (0.063)		-0.513*** (0.133)
22–28 days after			-0.246*** (0.073)		-0.654*** (0.149)
29–35 days after			-0.154** (0.077)		-0.371*** (0.138)
36–42 days after			-0.171** (0.081)		-0.335** (0.156)
43–49 days after			-0.182** (0.079)		-0.349** (0.159)
50–56 days after			-0.099 (0.077)		-0.364** (0.172)
57–63 days after			-0.066 (0.081)		-0.406** (0.166)
64–70 days after			-0.050 (0.083)		-0.336* (0.185)
> 70			0.048 (0.076)		-0.134 (0.180)
Number of trip-level observations	352253	352253	352253	187127	187127
Controls	No	Yes	Yes	No	Yes
Driver fixed effects	No	Yes	Yes	No	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	Yes	Yes	No	Yes

Notes: Identification of in-person coaching effects on driving performance. The time period under consideration is the period for which we have complete logs available from all coaches (01/01/2015–30/04/2016). Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, non-scheduled rides and having been coached.

***(*, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.13: In-Person Coaching Effects on Driving Performance: Braking and Cornering

Dependent variable:	Braking		Cornering	
	(1)	(2)	(3)	(4)
Post-coaching	-0.009 (0.026)	-0.016 (0.025)		-0.047 (0.038)
Day of first coaching			-0.109*** (0.034)	
1 –7 days after			-0.024 (0.025)	-0.190*** (0.039)
8 –14 days after			-0.063** (0.027)	-0.104*** (0.035)
15 –21 days after			-0.016 (0.025)	-0.027 (0.035)
22 –28 days after			-0.031 (0.026)	-0.079** (0.037)
29 –35 days after			-0.018 (0.028)	-0.098** (0.040)
36 –42 days after			-0.004 (0.029)	-0.054 (0.038)
43 –49 days after			0.017 (0.031)	-0.020 (0.045)
50 –56 days after			-0.007 (0.031)	-0.039 (0.047)
57 –63 days after			0.009 (0.030)	0.012 (0.051)
64 –70 days after			-0.005 (0.032)	-0.030 (0.051)
> 70			0.020 (0.037)	-0.015 (0.052)
Number of trip-level observations	187127	187127	187127	187127
Controls	No	Yes	Yes	Yes
Driver fixed effects	No	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	Yes	Yes	Yes

Notes: Identification of in-person coaching effects on driving performance. The time period under consideration is the period for which we have complete logs available from all coaches (01/01/2015–30/04/2016). Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. The dependent variables braking and cornering are measured as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. ***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

2.A.14 Temporal effects of in-person coaching at coach level

Figure 2.A.8: Temporal Effects of In-Person Coaching at Coach Level: Acceleration

Figure 2.A.9: Temporal Effects of In-Person Coaching at Coach Level: Braking

Figure 2.A.10: Temporal Effects of In-Person Coaching at Coach Level: Cornering

2.A.15 Change in ranking following coaching

Table 2.A.14: Change in ranking month ($m + 1$) vs. ($m - 1$) following coaching in month m

Dep. var	Rank month ($m + 1$)			
		Acceleration		
Coached month m		-0.006 (0.031)	-0.012 (0.035)	-0.074** (0.029)
Rank month ($m - 1$)		1.195*** (0.11)	0.697*** (0.122)	0.653*** (0.043)
obs.		754	736	1744
R^2		0.1765	0.0495	0.2018
Sample	Rank month ($m - 1$):	top-25%	25-50%	bottom-50%
		Braking		
Coached month m		0.004 (0.049)	-0.009 (0.039)	-0.068** (0.029)
Rank month ($m - 1$)		0.694*** (0.141)	0.527*** (0.136)	0.64*** (0.047)
obs.		751	731	1752
R^2		0.0392	0.0229	0.1576
Sample	Rank month ($m-1$):	top-25%	25-50%	bottom-50%
		Cornering		
Coached month m		-0.004 (0.021)	-0.024 (0.042)	-0.078*** (0.021)
Rank month ($m - 1$)		0.845*** (0.091)	0.758*** (0.094)	0.843*** (0.031)
obs.		740	739	1755
R^2		0.1478	0.0797	0.4318
Sample	Rank month ($m - 1$):	top-25%	25-50%	bottom-50%
		Fuel economy		
Coached month m		0.007 (0.088)	0.074 (0.055)	-0.087* (0.046)
Rank month ($m - 1$)		0.931*** (0.196)	0.451** (0.197)	0.565*** (0.062)
obs.		351	343	833
R^2		0.0607	0.0184	0.1304
Sample	Rank month ($m - 1$):	top-25%	25-50%	bottom-50%

Notes: All regressions include a constant. Standard errors clustered by driver.
 ***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

2.A.16 Complementarities between interventions

Figure 2.A.11: Treatment Level Effects In-Person Coaching: Fuel Economy and ABC

Notes: Driving performance in the 10 weeks before and after coaching based on trips with VDL and Intouro buses. The day of coaching itself is point 0 on the x -axis. The vertical spikes indicate 95% confidence intervals. The dependent variables braking and cornering are measured as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours and fill-in rides the interaction of bus type and date fixed effects. Coaches are excluded.

Table 2.A.15: Targeted peer-comparison Feedback Effects on Driving Performance - Driver NOT COACHED Receiving First Feedback

Dependent variable:	Fuel Economy			Acceleration		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-announcement	-0.172 (0.121)	-0.409*** (0.097)	-0.338*** (0.068)		-1.693*** (0.231)	-0.974*** (0.200)
T2 (1n/0p)	0.115 (0.232)	0.248 (0.220)			0.016 (0.420)	0.040 (0.426)
T3 (1n/1p)	0.350 (0.349)	0.357 (0.339)			0.470 (0.636)	0.442 (0.596)
T4 (3n/0p)	0.277 (0.329)	0.225 (0.297)			0.541 (0.524)	0.565 (0.530)
Post-feedback	-0.371*** (0.141)	-0.036 (0.127)	-0.023 (0.105)		-0.611** (0.289)	-0.539*** (0.252)
Post-feedback × T2 (1n/0p)	-0.426** (0.204)	-0.412*** (0.182)	-0.379** (0.155)	-0.337** (0.149)	-0.269 (0.386)	-0.079 (0.318)
Post-feedback × T3 (1n/1p)	0.001 (0.180)	-0.015 (0.171)	-0.117 (0.181)	-0.106 (0.180)	-0.370 (0.488)	-0.334 (0.426)
Post-feedback × T4 (3n/0p)	-0.354** (0.163)	-0.216 (0.140)	-0.304** (0.147)	-0.284** (0.141)	-0.889** (0.405)	-0.945*** (0.317)
Post-experiment	0.517*** (0.145)	0.519*** (0.091)	0.476*** (0.086)			0.161 (0.205)
Post-experiment × T2 (1n/0p)	-0.048 (0.207)	0.046 (0.127)	0.122 (0.130)	0.110 (0.115)	-0.026 (0.341)	0.024 (0.235)
Post-experiment × T3 (1n/1p)	0.144 (0.197)	0.131 (0.146)	0.084 (0.131)	0.091 (0.119)	0.158 (0.404)	0.193 (0.297)
Post-experiment × T4 (3n/0p)	-0.156 (0.199)	0.033 (0.143)	0.058 (0.134)	0.052 (0.125)	0.049 (0.387)	0.189 (0.316)
Constant	24.196*** (0.181)	21.809*** (0.207)	22.397*** (0.161)	24.067*** (0.386)	10.283*** (0.357)	8.538*** (0.361)
R ²	.0175	.409	.502	.526	.0451	.519
Number of trip-level observations	144305	136482	136482	136482	99562	94667
Controls	No	Yes	Yes	Yes	No	Yes
Weather dummies	No	Yes	Yes	No	No	Yes
Driver fixed effects	No	No	Yes	Yes	No	Yes
Day fixed effects	No	No	No	Yes	No	No
Bus type × day fixed effects	No	No	No	Yes	No	No

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours and fill-in rides. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.16: Targeted Peer-Comparison Feedback Effects on Driving Performance - Drivers COACHED Before First Feedback

Dependent variable:	Fuel Economy				Acceleration			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post-announcement	-0.452*** (0.091)	-0.392*** (0.105)	-0.455*** (0.072)		-2.617*** (0.180)	-1.069*** (0.194)	-1.261*** (0.168)	
T2 (1n/0p)	-0.552*** (0.266)	-0.518*** (0.194)			-0.510 (0.497)	-0.298 (0.404)		
T3 (1n/1p)	-0.332 (0.255)	-0.313 (0.207)			0.023 (0.582)	-0.046 (0.463)		
T4 (3n/0p)	-0.526*** (0.254)	-0.478*** (0.184)			-0.332 (0.479)	-0.155 (0.397)		
Post-feedback	-0.477*** (0.123)	0.007 (0.158)	-0.171* (0.100)		-0.436* (0.242)	-0.400 (0.291)	-0.673*** (0.249)	
Post-feedback × T2 (1n/0p)	0.090 (0.212)	0.071 (0.154)	0.117 (0.135)	0.120 (0.135)	0.404 (0.411)	0.349 (0.354)	0.260 (0.356)	0.238 (0.351)
Post-feedback × T3 (1n/1p)	0.165 (0.152)	0.134 (0.110)	0.082 (0.107)	0.069 (0.106)	0.004 (0.342)	0.185 (0.288)	0.183 (0.284)	0.200 (0.282)
Post-feedback × T4 (3n/0p)	0.173 (0.174)	0.126 (0.109)	0.064 (0.101)	0.035 (0.097)	0.059 (0.321)	0.286 (0.262)	0.242 (0.246)	0.254 (0.246)
Post-experiment	0.547*** (0.144)	0.486*** (0.093)	0.537*** (0.082)		0.308 (0.235)	0.117 (0.196)	0.220 (0.163)	
Post-experiment × T2 (1n/0p)	0.235 (0.216)	0.238 (0.169)	0.179 (0.149)	0.186 (0.140)	0.516 (0.352)	0.465 (0.321)	0.316 (0.273)	0.307 (0.266)
Post-experiment × T3 (1n/1p)	0.134 (0.185)	0.166 (0.127)	0.089 (0.105)	0.107 (0.107)	0.431 (0.296)	0.427* (0.249)	0.256 (0.216)	0.274 (0.216)
Post-experiment × T4 (3n/0p)	0.014 (0.171)	0.156 (0.114)	0.056 (0.104)	0.056 (0.105)	0.156 (0.313)	0.253 (0.251)	0.075 (0.193)	0.087 (0.192)
Constant	24.949*** (0.181)	22.243*** (0.142)	22.581*** (0.125)	23.768*** (0.289)	10.914*** (0.387)	8.389*** (0.315)	11.638*** (0.268)	11.793*** (0.377)
R ²	.0183	.424	.498	.517	.0458	.414	.561	.578
Number of trip-level observations	243265	232597	232597	232597	171092	164593	164593	164593
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Weather dummies	No	Yes	Yes	No	No	Yes	Yes	No
Driver fixed effects	No	No	No	Yes	No	No	Yes	Yes
Day fixed effects	No	No	No	Yes	No	No	No	Yes
Bus type × day fixed effects	No	No	No	Yes	No	No	No	Yes

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

***(*,); statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.17: Targeted Peer-Comparison Feedback Effects on Driving Performance - Drivers NOT COACHED Before First Feedback

Dependent variable:	Braking			Cornering				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post-announcement	-1.264*** (0.047)	-1.294*** (0.052)	-1.302*** (0.050)		-0.249*** (0.072)	-0.255*** (0.067)	-0.328*** (0.059)	
T2 (1n/0p)	0.046 (0.106)	0.065 (0.104)			0.197 (0.235)	0.195 (0.235)		
T3 (1n/1p)	-0.016 (0.125)	-0.004 (0.122)			-0.022 (0.245)	-0.038 (0.234)		
T4 (3n/0p)	0.203 (0.124)	0.222* (0.122)			0.392 (0.262)	0.391 (0.258)		
Post-feedback	0.182*** (0.051)	0.199*** (0.053)	0.199*** (0.057)		-0.115 (0.120)	-0.094 (0.115)	-0.098 (0.107)	
Post-feedback × T2 (1n/0p)	-0.081 (0.063)	-0.116* (0.068)	-0.130* (0.073)	-0.144* (0.078)	-0.085 (0.120)	-0.070 (0.118)	-0.047 (0.116)	-0.061 (0.118)
Post-feedback × T3 (1n/1p)	-0.085 (0.102)	-0.097 (0.099)	-0.088 (0.106)	-0.091 (0.103)	0.067 (0.131)	0.077 (0.121)	0.081 (0.119)	0.074 (0.120)
Post-feedback × T4 (3n/0p)	-0.280*** (0.090)	-0.291*** (0.091)	-0.282*** (0.090)	-0.279*** (0.092)	-0.318** (0.150)	-0.324** (0.148)	-0.226 (0.137)	-0.230 (0.139)
Post-experiment	0.006 (0.056)	-0.021 (0.048)	-0.003 (0.042)		-0.217*** (0.072)	-0.221*** (0.074)	-0.092* (0.051)	
Post-experiment × T2 (1n/0p)	0.066 (0.063)	0.034 (0.061)	0.026 (0.053)	0.040 (0.055)	0.191** (0.092)	0.172* (0.089)	0.059 (0.064)	0.046 (0.061)
Post-experiment × T3 (1n/1p)	0.104 (0.084)	0.132* (0.073)	0.099 (0.071)	0.113 (0.069)	0.117 (0.089)	0.113 (0.091)	0.002 (0.069)	-0.009 (0.068)
Post-experiment × T4 (3n/0p)	0.086 (0.068)	0.078 (0.061)	0.057 (0.060)	0.060 (0.057)	0.193* (0.111)	0.181 (0.112)	0.078 (0.095)	0.079 (0.093)
Constant	1.641*** (0.091)	1.362*** (0.103)	2.261*** (0.112)	1.991*** (0.174)	1.039*** (0.188)	1.162*** (0.202)	1.668*** (0.107)	1.615*** (0.133)
R ²	.0833	.196	.219	.29	.0306	.0973	.407	.421
Number of trip-level observations	99562	94667	94667	94667	100837	95938	95938	95938
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Weather dummies	No	Yes	Yes	No	No	Yes	Yes	No
Driver fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Day fixed effects	No	No	No	Yes	No	No	No	Yes
Bus type × day fixed effects	No	No	No	Yes	No	No	No	Yes

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours and fill-in rides. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

***(*, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.18: Targeted Peer-Comparison Feedback Effects on Driving Performance - Drivers COACHED Before First Feedback

Dependent variable:	Braking				Cornering			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post-announcement	-1.248*** (0.043)	-1.321*** (0.049)	-1.336*** (0.048)		-0.316*** (0.038)	-0.120* (0.069)	-0.225*** (0.051)	
T2 (1n/0p)	-0.124 (0.091)	-0.111 (0.086)			-0.196 (0.191)	-0.172 (0.176)		
T3 (1n/1p)	-0.145 (0.092)	-0.138 (0.089)			-0.228 (0.175)	-0.235 (0.159)		
T4 (3n/0p)	-0.167 (0.106)	-0.168* (0.097)			-0.167 (0.169)	-0.125 (0.166)		
Post-feedback	0.017 (0.048)	-0.156*** (0.051)	-0.159*** (0.051)		-0.122** (0.052)	-0.056 (0.086)	-0.157*** (0.066)	
Post-feedback × T2 (1n/0p)	0.135* (0.076)	0.112 (0.071)	0.101 (0.071)	0.087 (0.067)	0.029 (0.115)	0.026 (0.109)	0.061 (0.111)	0.054 (0.110)
Post-feedback × T3 (1n/1p)	0.069 (0.073)	0.055 (0.068)	0.056 (0.068)	0.055 (0.066)	0.052 (0.090)	0.065 (0.085)	0.075 (0.090)	0.070 (0.088)
Post-feedback × T4 (3n/0p)	0.113 (0.082)	0.122 (0.079)	0.095 (0.076)	0.097 (0.074)	0.149** (0.074)	0.148** (0.073)	0.139* (0.075)	0.130* (0.074)
Post-experiment	0.134** (0.053)	0.060 (0.044)	0.066 (0.045)		-0.066 (0.064)	-0.092 (0.056)	-0.026 (0.035)	
Post-experiment × T2 (1n/0p)	0.007 (0.071)	-0.011 (0.063)	-0.008 (0.062)	-0.017 (0.061)	0.047 (0.081)	0.067 (0.071)	-0.021 (0.054)	-0.024 (0.054)
Post-experiment × T3 (1n/1p)	0.036 (0.071)	0.055 (0.053)	0.039 (0.053)	0.029 (0.054)	0.014 (0.071)	0.032 (0.058)	-0.033 (0.042)	-0.035 (0.042)
Post-experiment × T4 (3n/0p)	0.030 (0.065)	0.068 (0.064)	0.055 (0.062)	0.042 (0.059)	0.052 (0.081)	0.067 (0.069)	-0.039 (0.051)	-0.040 (0.050)
Constant	1.786*** (0.079)	1.392*** (0.083)	2.264*** (0.093)	2.113*** (0.119)	1.178*** (0.131)	1.304*** (0.134)	1.492*** (0.075)	1.643*** (0.117)
R ²	.0767	.206	.233	.291	.0235	.101	.379	.391
Number of trip-level observations	171092	16453	16453	16453	174143	167632	167632	167632
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Weather dummies	No	Yes	Yes	No	No	Yes	Yes	No
Driver fixed effects	No	No	No	Yes	No	No	Yes	Yes
Day fixed effects	No	No	No	Yes	No	No	No	Yes
Bus type × day fixed effects	No	No	No	Yes	No	No	No	Yes

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. Treatments vary in the number of positive and negative peer-comparison messages on the comfort driving dimensions (acceleration, braking, cornering). Messages are targeted in the sense that they are only provided if a driver performs relatively poor (bottom 50%) or good (top 25%) compared to a reference group of colleagues. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report.

***(*, *); statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.19: Targeted Peer-Comparison Feedback Effects on Driving Performance [Groups: Age]

Dependent variable:	Fuel economy	Acceleration	Braking	Cornering
	(1)	(2)	(3)	(4)
Post-announcement=1	-0.371*** (0.065)	-0.911*** (0.185)	-1.214*** (0.049)	-0.287*** (0.054)
Post-announcement=1 × agegroup=1	-0.310*** (0.114)	-0.800*** (0.280)	-0.093 (0.069)	-0.082 (0.087)
Post-announcement=1 × agegroup=2	0.016 (0.105)	0.191 (0.251)	-0.135* (0.081)	-0.063 (0.075)
Post-announcement=1 × agegroup=4	0.047 (0.092)	-0.011 (0.230)	-0.023 (0.071)	-0.074 (0.075)
Post-feedback=1	-0.132** (0.065)	-0.972*** (0.172)	0.008 (0.028)	-0.210*** (0.044)
Post-feedback=1 × agegroup=1	-0.095 (0.112)	0.113 (0.270)	-0.026 (0.040)	0.062 (0.057)
Post-feedback=1 × agegroup=2	-0.073 (0.125)	0.049 (0.300)	-0.027 (0.046)	0.043 (0.083)
Post-feedback=1 × agegroup=4	0.069 (0.103)	0.391* (0.213)	0.030 (0.044)	0.167*** (0.054)
Post-experiment=1	0.599*** (0.043)	0.436*** (0.106)	0.070*** (0.024)	-0.024 (0.026)
Post-experiment=1 × agegroup=1	-0.046 (0.076)	-0.244 (0.164)	-0.077** (0.035)	-0.026 (0.035)
Post-experiment=1 × agegroup=2	-0.033 (0.078)	-0.137 (0.183)	-0.010 (0.037)	-0.071 (0.053)
Post-experiment=1 × agegroup=4	0.014 (0.067)	-0.103 (0.174)	0.003 (0.034)	-0.067* (0.035)
Constant	22.555*** (0.078)	11.357*** (0.188)	2.314*** (0.059)	1.642*** (0.052)
R ²	.531	.45	.198	.396
# trip-level observations	484918	349879	349879	349879
Controls	Yes	Yes	Yes	Yes
Weather dummies	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes
Day fixed effects	No	No	No	No
Bus type × day fixed effects	No	No	No	No

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report. Default agegroup=3 (55-59 years); agegroup=1: < 50 years; agegroup=2: 50 – 54 years; agegroup=4: ≥ 60 years.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.20: Targeted Peer-Comparison Feedback Effects on Driving Performance [Groups: Years of Service]

Dependent variable:	Fuel economy	Acceleration	Braking	Cornering
	(1)	(2)	(3)	(4)
Post-announcement=1	-0.490*** (0.075)	-1.193*** (0.182)	-1.254*** (0.042)	-0.342*** (0.056)
Post-announcement=1 $\times$ tenuregroup=1	0.017 (0.101)	0.082 (0.245)	-0.016 (0.070)	0.014 (0.076)
Post-announcement=1 $\times$ tenuregroup=2	0.016 (0.120)	0.073 (0.280)	-0.099 (0.070)	-0.008 (0.092)
Post-announcement=1 $\times$ tenuregroup=4	0.180* (0.106)	0.304 (0.257)	0.051 (0.068)	0.016 (0.076)
Post-feedback=1	-0.155* (0.090)	-0.938*** (0.193)	-0.008 (0.029)	-0.130*** (0.043)
Post-feedback=1 $\times$ tenuregroup=1	-0.082 (0.117)	0.041 (0.273)	0.056 (0.042)	-0.006 (0.064)
Post-feedback=1 $\times$ tenuregroup=2	0.006 (0.129)	0.227 (0.247)	-0.026 (0.046)	-0.066 (0.062)
Post-feedback=1 $\times$ tenuregroup=4	0.129 (0.110)	0.242 (0.248)	0.016 (0.045)	0.012 (0.058)
Post-experiment=1	0.637*** (0.054)	0.412*** (0.128)	0.049* (0.025)	-0.026 (0.033)
Post-experiment=1 $\times$ tenuregroup=1	-0.055 (0.078)	-0.068 (0.178)	-0.008 (0.035)	-0.042 (0.042)
Post-experiment=1 $\times$ tenuregroup=2	-0.073 (0.084)	-0.158 (0.190)	-0.022 (0.039)	-0.059 (0.044)
Post-experiment=1 $\times$ tenuregroup=4	-0.084 (0.070)	-0.178 (0.179)	0.019 (0.038)	-0.043 (0.043)
Constant	22.562*** (0.080)	11.359*** (0.190)	2.318*** (0.058)	1.636*** (0.053)
R ²	.531	.449	.198	.396
# trip-level observations	484918	349879	349879	349879
Controls	Yes	Yes	Yes	Yes
Weather dummies	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes
Day fixed effects	No	No	No	No
Bus type $\times$ day fixed effects	No	No	No	No

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report. Default tenuregroup=3 (16-29 years); tenuregroup=1: < 8 years; tenuregroup=2: 8 – 15 years; tenuregroup=4: $\geq$ 30 years.

***(*, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 2.A.21: Targeted Peer-Comparison Feedback Effects on Driving Performance [Groups: Gender]

Dependent variable:	Fuel economy	Acceleration	Braking	Cornering
	(1)	(2)	(3)	(4)
Post-announcement=1	-0.429*** (0.044)	-1.100*** (0.104)	-1.269*** (0.028)	-0.341*** (0.032)
Post-announcement=1 × gendergroup=1	-0.063 (0.090)	0.155 (0.266)	0.032 (0.098)	0.038 (0.086)
Post-feedback=1	-0.123** (0.049)	-0.750*** (0.103)	0.008 (0.018)	-0.134*** (0.028)
Post-feedback=1 × gendergroup=1	-0.199 (0.154)	-0.625 (0.397)	-0.044 (0.045)	-0.058 (0.068)
Post-experiment=1	0.612*** (0.028)	0.369*** (0.066)	0.055*** (0.013)	-0.059*** (0.015)
Post-experiment=1 × gendergroup=1	-0.298** (0.119)	-0.580** (0.229)	-0.094* (0.049)	-0.032 (0.044)
Constant	22.554*** (0.079)	11.368*** (0.193)	2.317*** (0.059)	1.637*** (0.054)
R ²	.529	.448	.197	.393
Number of trip-level observations	484918	349879	349879	349879
Controls	Yes	Yes	Yes	Yes
Weather dummies	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes
Day fixed effects	No	No	No	No
Bus type × day fixed effects	No	No	No	No

Notes: Identification of the treatment effects on driving performance. The time period under consideration is from 01/01/2015 until 31/01/2017. The post-announcement dummy variable is one in the period from 09/11/2015 onwards (feedback announcement), zero otherwise. Drivers are considered to be in the post-feedback period when they have received at least one report in the past. For most drivers, this was at 15/12/2015 and after. The post-experimental period starts at 15/11/2016 when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. The dependent variable fuel economy is measured in liters/100km and acceleration as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, fill-in rides and having been coached. Weather dummies: daily temperature, wind and rainfall. A no-report indicator is included to capture drivers operating after 15/12/2015 (first feedback round) but who have not yet received their first report. Default gendergroup=0 (males).

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Chapter 3

Learning from high-achieving peers: A quasi-experiment in the field*

Abstract

Do workers improve their productivity if they are coached by high-achieving peers? This chapter presents the results from a quasi-experimental evaluation of a peer-based coaching program at a large public transport company. The program involves the provision of tailored on-the-road coaching sessions to bus drivers by experienced peers. Coach logs pin down for every driver when exactly they were coached. Using close to two million trip-level observations from 1,032 bus drivers over a two-and-a-half year sample period, I find that coaching by peers improves the performance of drivers on multiple productivity dimensions. Improvements are concentrated among drivers in the bottom half of the performance distribution. The positive effects diminish over time for aggregate productivity (fuel efficiency), but appear more persistent for disaggregate productivities (acceleration and braking), suggesting some habituation in these dimensions. These results broadly replicate the findings in Romensen and Soetevent (2020a) and extend it to workplace settings with no parallel peer-comparison feedback in place.

*This chapter is based on Romensen (2020a).

3.1 Introduction

A firm's potential for learning and productivity growth may be hampered by the fragmentation and unequal distribution of knowledge and skills among its employee base. One promising and inexpensive way of addressing this concern is to facilitate as a firm exchanges between workers and high-achieving peers. While learning from others is long recognized as an important input to individual learning mechanisms (Arrow 1994), it remains surprisingly unclear how initiatives by firms that facilitate such exchanges affect workers' subsequent performance. Previous studies suggest that social interactions in the workplace improve workers' productivity mainly through peer pressure (Mas and Moretti 2009, Falk and Ichino 2006). Less research is available that focuses on productivity gains due to spillovers of knowledge and skills. How workers adapt to direct teaching from high-achieving colleagues is therefore an important open question. In particular because it may carry implications that are not confined to the workforce but readily extend to many other policy-relevant domains.

In this chapter, I perform a quasi-experimental evaluation of a peer-based coaching program at a large public transport company. In the program, experienced bus drivers are designated as coaches and provide tailored feedback to colleagues during on-the-road coaching sessions. Detailed coach logs allow me to uniquely identify all driver-specific coaching moments. The short-run and long-run impact of a single coaching session on worker productivity is evaluated using close to two million trip-level observations from 1,032 tenured bus drivers in a two-and-a-half year period.

The coaching program and the public transport setting are ideal for the identification and evaluation of skill and knowledge transfers between workers. First, by design we know the direction of the transfer and when the interaction took place. Second, the nature of the job is such that there is only very limited scope for other forms of learning from colleagues (e.g., observational learning). Also, bus driving is an individual task that is precisely measured and independent from peers' performance. Third, the job itself is not too easy nor too complicated and there is therefore ample scope for learning. Finally, on-board computers in buses generate rich

and high-frequency productivity data, as measured by various measures of driving behavior, and the short- and long run impact of the coaching program can therefore be precisely estimated.

In a companion study, Romensen and Soetevent (2020a) evaluate the same coaching program in a different concession area of the public transport company, alongside a targeted peer-comparison feedback program. They find positive effects of coaching on multiple productivity dimensions that are persistent for up to about eight weeks. The positive effects are mainly observed among drivers in the bottom half of the performance distribution. Further analyses show that the parallel peer-comparison feedback program does not seem to interfere with the effectiveness of coaching.

The contribution of the present chapter is twofold. First, I provide precise empirical estimates of the short-run and long-run causal impact of peer-based coaching on individual worker productivity. The impact is identified in a workplace setting without a targeted peer-comparison feedback program in place, and thereby extends the earlier findings from Romensen and Soetevent (2020a). Second, I perform a scientific replication of the main findings in Romensen and Soetevent (2020a). Hamermesh (2007) defines a scientific replication as replicating research findings with a different sample and a similar but possibly not identical model. As digital technologies, like the on-board computers in buses, significantly reduce replication costs (Goldfarb and Tucker 2019), performing scientific replications to validate earlier findings has become much more feasible. Several researchers have recently called for more replications in economics research (Mueller-Langer, Fecher, Harhoff, and Wagner 2019, Coffman, Niederle, and Wilson 2017).

I find that the estimated treatment effects of on-the-road coaching by experienced peers in the new sample area are qualitatively very similar to the effects reported in Romensen and Soetevent (2020a). On the day of coaching, bus drivers drive more efficiently and need about 0.6 liters of fuel (0.58 SD, $p < 0.001$) less to drive 100 kilometers. A lower fuel need is also observed in the first six weeks following coaching, but after that no difference is found with non-coached drivers. Drivers therefore do not seem to internalize the more efficient driving style. The pattern of eventual decay in treatment effects is less pronounced for acceleration

and braking, although effect sizes appear to become smaller over time. As in Romensen and Soetevent (2020a), most of the gains from coaching are concentrated among the drivers with relatively low baseline performance scores. Improvements from coaching are observed irrespective of whether drivers received a performance feedback report in the past.

The findings in this study shed new light on peer effects in the workplace. Existing research in this context has documented that social pressure serves as the main channel through which peer effects operate (Mas and Moretti 2009, Falk and Ichino 2006).¹ In these studies, however, workers perform an easy task with few learning opportunities, and can readily observe coworkers' productivity to infer an acceptable pace of working. In the setting of this study, social pressure is unlikely to drive the observed peer effects. While it may partially account for the improvement in driving behavior during the coaching session, it cannot explain why improvements persist in the ensuing weeks. I therefore favor an interpretation where workers, at least temporarily, acquire new skills and knowledge through learning from peers, in line with the original objective of the coaching program.

A few other studies, mainly in the areas of research and education, have examined productivity gains due to knowledge spillovers between peers. Waldinger (2012) finds no evidence of peer effects among university scientists. Azoulay, Graff Zevin, and Wang (2010) find that the productivity of researchers decreases when a highly productive collaborator ("superstar") dies prematurely and unexpectedly. Jackson and Bruegmann (2009) conclude that student outcomes improve if teachers have more effective colleagues. In a field experiment, Papay et al. (2020) also find that student outcomes improve if low-skilled teachers are paired with high-skilled peers for coaching purposes. Outside of schools and universities, Lindquist, Sauermann, and Zenou (2017) report evidence of knowledge spillovers among call center employees. Chan, Li, and Pierce (2014) consider salespersons of a department store and suggest that peer-based learn-

¹Peer pressure may in part also explain the increases in productivity found among data entry workers (Kaur, Kremer, and Mullainathan 2010), soft-fruit pickers (Bandiera, Barankay, and Rasul 2010), call center employees (Lindquist, Sauermann, and Zenou 2017), and salespersons of a department store (Chan, Li, and Pierce 2014). See Cornelissen, Dustmann, and Schönberg (2017) for a further discussion.

ing among these salespersons are based on transfers of skills and knowledge. Sandvik, Saouma, Seegert, and Stanton (2020) also consider a setting with salespersons and observe persistent productivity improvements from peer learning. The findings presented in this chapter are broadly in line with these previous studies, but also suggest that productivity gains can wane and that additional interventions may be needed to sustain improvements.

The results of this study more generally contribute to the recent literature on non-financial interventions to improve productivity and conservation efforts in the workplace. Pay-for-performance schemes to trigger behavioral changes are often not feasible in workplace settings because of reluctant work councils and unions. In such cases, non-financial interventions may be a viable and promising alternative. Blader, Gartenberg, and Prat (2020) examine the impact of relative performance feedback on truck driver performance and find that the feedback increases (decreases) fuel efficiency in plants without (with) a culture of teamwork and empowerment. Gosnell, List, and Metcalfe (2020) study the effect of various behavioral interventions on the fuel efficiency of airline captains. They find that the interventions (especially performance monitoring and target setting) lead to significant improvements on various fuel-saving dimensions. This chapter suggests another promising channel through which firms can improve the productivity and conservation efforts of workers, namely peer coaching. New technologies and the digitalization of worker performance records make it easier for firms to identify and pair low-skilled workers and high-skilled workers and thus to facilitate peer learning. This opens up many new opportunities for designing and evaluating peer coaching programs in the workplace.

For firms and policy makers interested in improving productivity and conservation efforts, the present research suggests that it can be beneficial to employ interventions that target workers directly. Programs that facilitate interactions between workers and high-achieving colleagues can induce the transfers of skills and knowledge that are needed to foster positive changes in worker behavior. Even when the improvements at the worker level are relatively small, the aggregate impact on economic and environmental outcomes may very well be large and meaningful.

This chapter proceeds as follows. Section 3.2 details the field setting of this study. Section 3.3 provides further information on the peer coaching program that is being evaluated. Section 3.4 elaborates on the data and the phase in of the coaching program in the sample area. Section 3.5 presents the analyses and the results. Section 3.6 concludes the chapter.

3.2 The setting

This study is a continuation of the collaboration started with Arriva Netherlands (henceforth Arriva) in Romensen and Soetevent (2020a). Arriva is a subsidiary of Deutsche Bahn AG and operates various transport modes in the Dutch public transport industry. This study focuses on two concession areas in the southwestern part of the Netherlands. The sample area overlaps with the Randstad metropolitan area, as shown in Figure 3.A.1, and has about 140 unique routes (bus lines) in total. There are eight base locations and each of the 1,232 drivers operating in the area belongs to one of these locations. The period under consideration in this study is from 1 July 2015 until 31 December 2017. During this period, Arriva transported about 51 million passengers, based on the raw trip-level data with check-ins from public transport cards.

The sample period encompasses the complete implementation process of a campaign aimed at economical and comfortable driving behavior. This EcoManager campaign started in the concession area discussed in Romensen and Soetevent (2020a) and is now implemented in the concession areas discussed in this chapter. As part of the campaign, electronic on-board recorders (EOBRs) are installed in the entire fleet of buses. The EOBR complements the existing technology in the bus, which only measures fuel economy, by measuring the performance on other driving dimensions as well, such as acceleration, braking, and cornering (ABC). It does so by registering so-called “events” whenever an action by a driver is in excess of pre-specified performance threshold levels. Scoring an acceleration event, for example, means that the driver accelerated too fast. Taken together, the on-board technology provides a complete picture of driver performance, capturing aggregate (fuel economy) and disaggregate (ABC) dimensions.

Figure 3.1: Timeline Study

Figure 3.1 shows the timeline of the campaign. The EOBR installation was completed by the end of August 2016, after which the EOBRs started communicating ABC performance data per 1 September 2016. The official introduction to the bus drivers was during a kickoff event on 1 December 2016. At this event the workings of the EOBR were explained and the on-the-road coaching program was introduced. It was also communicated to the drivers that the data from the EOBR will not be used in formal evaluations, such that we can rule out career considerations as the source of changes in driving behavior. After the kickoff event, coaches started with their coaching sessions.

Two other initiatives were introduced during the kickoff event. The first initiative is a monthly driver-specific feedback report with a succinct summary of driving performance during the previous month. The first report was distributed around 15 January 2017. This feedback report is akin to the report discussed in Romensen and Soetevent (2020a) with the exception that it does not contain targeted peer-comparison feedback. The other initiative is a LED array that is installed in all buses. The LED arrays provide drivers with feedback on the ABC dimensions during the ride. LED arrays were also part of the campaign in Romensen and Soetevent (2020a). While these two initiatives enhance performance saliency to all drivers, they do little in terms of guiding drivers on how to actually improve driving behavior, which is the added value of the coaching program.

3.3 The in-person coaching program

The coaching program consists of seven experienced bus drivers, who were recruited as coaches to provide colleagues with feedback during on-the-road sessions. The coaches are on average about 53 years old, have more than 20 years experience as a bus driver, and are all male. Prior to the start of the program, all coaches participated in four training sessions. Two sessions were taught in-class by an external agency and dedicated to making coaches familiar with all aspects of economical driving behavior. The other sessions were practice sessions in the field. Coaches received tips on how to approach drivers and to communicate feedback.

A detailed discussion of the coaching program and the structure of coaching sessions is provided in Romensen and Soetevent (2020a). In sum, during a session an experienced peer rides along with a bus driver and afterwards provides tailored feedback to the driver based on his or her performance during the observation. The session is not announced to the driver beforehand. The feedback is solely based on driver performance during the session because coaches were not granted access to additional performance data of drivers, nor did the company communicate to the coaches which drivers perform well or poor. This feature of the program is consistent with a quasi-random order of coaching, as will be discussed in more detail in Section 3.4.

For the first three weeks after the campaign kickoff event, coaches were freed up from other obligations in order to have extra time available for coaching activities. After these three weeks, coaches were granted one day a week for riding along with drivers, and they could freely choose this day. Coaches were instructed to meticulously log all of their activities. I use these logs to identify the coaching moment for every driver in the sample.

3.4 Data and phase-in of the coaching program

The initial dataset covers all trips in the sample area during the period from 1 July 2015 till 31 December 2017. To facilitate the comparison of results with Romensen and Soetevent (2020a), I will restrict the attention

to tenured drivers (84% of all drivers). After performing a few cleaning steps, detailed in Table 3.A.1, the dataset used for the analyses contains 1,969,408 trip-level observations. The on-board computers make it possible to match these observations to the 1,032 tenured bus drivers operating in the area. The dataset is further complemented with weather data, coach logs, and administrative databases containing information on drivers, trips, and check-ins with public transport cards.²

Figure 3.A.2 shows the moment of first coaching for the drivers in the sample. Most drivers participate in their first coaching session in the weeks after the kickoff event on 1 December 2016. Coaching intensity gradually decreases and remains at the same level after the first feedback report in mid January 2017. After mid 2017, first coaching moments become scarce and the share of coached drivers remains more or less the same. In total, the logs indicate that 595 drivers (58% of all drivers) participated in at least one coaching session during the sample period. Some drivers (88 in total) participated in additional coaching sessions, mainly in the second quarter of 2017. These sessions were meant as a check by coaches to see whether drivers habituated to the economical driving style. I control for these extra sessions in the analyses.

Further inspection reveals that the assignment of coaching sessions follows a quasi-random pattern. Drivers appear not to be selected based on observables in the data. Table 3.1 shows the results for baseline performance and non-performance characteristics by comparing coached drivers and not-yet-coached drivers who operate from the same base location on dates when coaching took place at that location.³ If the order of coaching is quasi-random, there should not be a systematic difference between these two groups. The table reports the unadjusted p -values of balance tests as well as p -values that are adjusted for multiple hypothesis testing (MHT).⁴ Looking at fuel economy and the ABC dimensions, I find that for

²Weather data are collected from the Royal Netherlands Meteorological Institute (KNMI). These data include for each date in the sample period information about the average temperature, rainfall, and wind speed in the sample area.

³Table 3.1 only considers the six base locations where coaches are stationed. Coaching activity was too limited to make meaningful comparisons at the two base locations where no coaches are stationed (less than 1% of the drivers received coaching).

⁴The Bonferroni and Holm methods are used to adjust the p -values for multiple hy-

Table 3.1: Quasi-Random Coaching: Balancing Tests on Baseline Performance and Non-Performance Descriptives

	C	NC	$\Delta(\text{C-NC})$	stand. p -value	Bonf. corr.	Holm corr.
<i>Baseline performance</i>						
Fuel economy	27.47	27.52	-0.05	0.6292	1.0000	1.0000
Acceleration	13.15	13.29	-0.14	0.5546	1.0000	1.0000
Braking	2.55	2.49	0.06	0.3435	1.0000	1.0000
Cornering	1.36	1.39	-0.02	0.7571	1.0000	1.0000
<i>Demographics</i>						
Year of birth	1966.89	1966.25	0.64	0.4163	1.0000	1.0000
Year of employment	2001.12	2001.61	-0.49	0.5185	1.0000	1.0000
Share of FTE ≥ 0.9	0.54	0.55	-0.01	0.8432	1.0000	1.0000
Share of female drivers	0.19	0.18	0.02	0.7452	1.0000	1.0000
<i>Trip-specific variables</i>						
Punctuality	-3.08	-3.08	0.01	0.8838	1.0000	1.0000
Distance traveled (in km.)	20.57	22.03	-1.46	0.0000***	0.0000***	0.0000***
Number of passengers	19.04	20.15	-1.11	0.0118**	0.2006	0.0354**
Number of bus stops	30.94	33.55	-2.61	0.0000***	0.0000***	0.0000***
Morning rush hours	0.26	0.18	0.08	0.1363	1.0000	0.6815
Evening rush hours	0.12	0.23	-0.10	0.0278	0.4726	0.1112
<i>Share of rides on bus types</i>						
VDL	0.60	0.57	0.03	0.6397	1.0000	1.0000
DAF	0.12	0.12	0.01	0.9005	1.0000	1.0000
Volvo	0.28	0.31	-0.03	0.5516	1.0000	1.0000

Notes: Unit of observation is the driver. For every coaching date, the mean baseline performance and non-performance related variables of drivers who receive their first coaching (C) is compared to that of not-yet-coached colleagues (NC). Only the six base locations where coaches are stationed are considered. Reported are the mean values over all coaching dates. Standard p -values as well as p -values that use a Bonferroni and Holm correction for multiple hypothesis testing are reported. Fuel economy in liters/100km; Performance on the ABC dimensions as the number of events/10km. Fewer events means better driving behavior. Punctuality is the difference in minutes between actual and planned driving time. Number of passengers is based on check-ins with public transport cards. Morning and evening rush hours are from 7:00-10:00 and 16:00-19:00, respectively.

***(**, *): the corresponding p -values are less than 1% (5% or 10%).

all locations combined there is indeed no significant difference in baseline performance. Similar results are obtained when considering demographics and trip characteristics. For two trip-specific variables, distance traveled and the number of bus stops, the difference between coached and not-yet-coached drivers is statistically significant, albeit small in absolute terms.⁵ In the analyses, discussed in Section 3.5, the preferred specification of the regression model will include trip-specific controls as well as route fixed effects to account for this difference.

The drivers in the sample are, on average, about 50 years old and have more than 15 years of experience as a bus driver. Most drivers are male (82%). About 82% of the drivers operate on a full time basis and have ≥ 0.8 FTE contracts. Buses on an average trip carry about 18 passengers and travel for 20 kilometers, with 30 bus stops along the way. Bus trips in the concession areas are typically inter-city trips (87%). The demographics and trip characteristics in this study and the study reported in Romensen and Soetevent (2020a) look very much alike.

Comparing columns (4) and (5) in Table 3.2 gives an indication of the influence of driver performance on the variation in fuel economy, and thus of the scope for improvement in driving behavior.⁶ In the table, I examine which part of the variation in fuel economy can be attributed to differences in bus types (column (1)), weather (column (2)), and trip-specific characteristics (column (3)). In column (4), I replace the weather dummies with day fixed effects. The estimates in the table suggest that the influence of weather on fuel economy is limited. Driving during evening rush hours, carrying more passengers, encountering more bus stops, and driving behind schedule are all trip-specific factors that worsen fuel economy. After

pothesis testing. With the Bonferroni method the unadjusted p -values are multiplied with the number of hypotheses that are tested. In the Holm method, the unadjusted p -values are first ranked from largest to smallest and afterwards each unadjusted p -value is multiplied with its rank.

⁵The length of the average trip of coached drivers (20.57 kilometers) is slightly shorter than that of not-yet-coached drivers (22.03 kilometers). Coached drivers also encounter slightly fewer bus stops on an average trip (30.94 stops) than not-yet-coached drivers (33.55 stops). Given that the differences are small in absolute terms, it can reasonably be argued that these two trip components did not influence which drivers were ultimately exposed to coaching.

⁶The tables for the ABC dimensions can be found in Appendix section 3.A.4.

Table 3.2: Determinants of Fuel Economy

Dependent variable:	Fuel economy				
	(1)	(2)	(3)	(4)	(5)
DAF	2.281*** (0.317)	2.299*** (0.317)	1.626*** (0.288)	1.601*** (0.284)	1.385*** (0.284)
VDL 10m	1.561*** (0.162)	1.572*** (0.162)	2.034*** (0.122)	2.008*** (0.118)	2.058*** (0.088)
Volvo 13m	4.488*** (0.133)	4.499*** (0.132)	4.567*** (0.101)	4.548*** (0.098)	4.468*** (0.093)
Volvo 14m	8.253*** (0.157)	8.229*** (0.156)	7.998*** (0.118)	7.918*** (0.117)	7.948*** (0.116)
Volvo hybrid	8.254*** (0.339)	8.315*** (0.338)	7.357*** (0.310)	7.318*** (0.306)	7.082*** (0.303)
Rush hour 7-10am			-0.078*** (0.026)	-0.092*** (0.025)	-0.080*** (0.017)
Rush hour 4-7pm			1.025*** (0.031)	1.052*** (0.031)	1.067*** (0.020)
Non-scheduled trip			0.005 (0.146)	0.008 (0.142)	-0.065 (0.066)
No. of stops per km.			1.993*** (0.074)	2.343*** (0.072)	2.365*** (0.068)
Trip length (in km.)			-0.114*** (0.003)	-0.108*** (0.003)	-0.104*** (0.002)
Ln(No. of passengers)			2.258*** (0.023)	2.194*** (0.023)	2.136*** (0.021)
Punctuality			0.101*** (0.005)	0.098*** (0.005)	0.153*** (0.002)
Constant	25.602*** (0.094)	25.123*** (0.095)	18.550*** (0.156)	15.968*** (0.362)	17.052*** (0.541)
R ²	.342	.351	.457	.465	.521
Number of trip-level observations	1783282	1783282	1783282	1783282	1783282
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Fuel economy in liters/100km. Default categories: trips outside rush hours completed with a VDL 12m bus. Trip punctuality is the difference in minutes between actual and scheduled driving time. Standard errors are clustered by driver.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 3.3: Driver Variation in Fuel Economy and ABC Comfort Outcomes

Driving dimension	μ	σ_r	c_v	$\Delta(p_{90} - p_{10})$	Residual correlation		
					Acc.	Braking	Corn.
Fuel economy	27.39	1.45	0.05	3.56	0.288	0.246	0.178
Acceleration	8.84	2.26	0.26	5.00		0.192	0.139
Braking	2.15	1.11	0.52	2.35			0.260
Cornering	1.95	1.23	0.63	2.50			

Notes: Fuel economy: liters/100km. ABC dimensions: no. of events/10km. σ_r : residual standard deviation; c_v : coefficient of variation. $\Delta(p_{90} - p_{10})$: difference in performance 90th vs. 10th percentile driver.

controlling for all time-varying variables, structural differences in driver performance explain an additional 5.6 percentage points of the variation in fuel economy (column (5)). To put this in perspective, interventions that can move a driver’s average fuel economy from the 90th percentile to the 10th percentile result in an economically significant reduction in fuel consumption of 3.56 liters per 100km ($\approx 13\%$). The scope for improvement is even higher on the ABC dimensions, as measured by the coefficient of variation in Table 3.3, although it should be noted that the upward potential in terms of braking and cornering is bounded by the already small average number of events per 10km. The presence of a scope for improvement on all driving dimensions suggest that if drivers benefit from in-person coaching by peers, then this can have sizable benefits in terms of both more economical and comfortable driving behavior.

3.5 Results

Coached drivers are observed in the trip-level data before and after they participated in a coaching session. In the same time frame the driving performance of not-yet-coached drivers is also observed. A differences-in-differences (DID) approach is therefore adopted to identify the causal impact of a single coaching session on the driving performance of bus drivers. Given the data similarities and to ease the comparison of results, I will estimate the same DID regression as in Romensen and Soetevent (2020a):

$$\begin{aligned}
 Y_{its} = & \delta_0 I\{t = t_i^c\} + \sum_{\tau=1}^{10} \delta_{\tau} I\{t - t_i^c \in (7(\tau - 1), 7\tau]\} + \delta_{10}^+ I\{t - t_i^c > 70\} \\
 & + \sum_{\tau=1}^{10} \gamma_{-\tau} I\{t - t_i^c \in [-7\tau, -7(\tau - 1))\} \\
 & + X_{its} \cdot \theta + \mu_i + \kappa_b + v_t + \zeta_{bt} + \xi_r + \epsilon_{its}.
 \end{aligned} \tag{3.1}$$

The dependent variable Y_{its} captures the outcome variable (fuel economy or ABC) per driver (i), day (t), and bus trip (s). Fuel economy is measured as the number of liters of fuel needed to drive 100 kilometers. Performance

on the ABC dimensions is quantified as the number of events per 10 kilometers (scoring fewer events indicates better performance). The day t_i^C captures the day at which driver i participated in a coaching session for the first time. The date of t_i^C may differ per driver because the coaching program was phased in over time. The regression specification identifies the effect of coaching in the ten weeks preceding and following the coaching session. The coefficient δ_{10}^+ absorbs any remaining effect after ten weeks. The period more than ten weeks before coaching is the reference period. The vector X_{its} contains controls for additional coaching sessions, the number of passengers and bus stops, morning and evening rush hours, travel distance, fill-in rides, and the specific bus model used.⁷ The specification also includes dummies that account for driver (μ_i), bus type (κ_b), day (v_t), and route (ξ_r) fixed effects.⁸ To account for within-driver correlation patterns in the error term (Bertrand, Duflo, and Mullainathan 2004), I will estimate the model with robust standard errors that are clustered at driver level.

Coefficient plots are presented in Figure 3.2. The plots show the effects of coaching on fuel economy and the ABC dimensions in the ten weeks prior and following the day of coaching. The results are very much in line with those found in Romensen and Soetevent (2020a). On the day of the coaching session, fuel economy improves significantly: the fuel amount needed to drive 100 kilometers reduces with about 0.6 liters ($0.41\sigma_r$). The effects of coaching diminish over time and after about six weeks become statistically indistinguishable from zero. This suggests that drivers did not habituate to the more efficient driving style. Figure 3.2 also shows that in the weeks prior to coaching, there are no significant differences between coached and not-yet-coached drivers in terms of fuel economy.

In Table 3.4, I summarize the results from regressions that take the entire period preceding coaching as the baseline period.⁹ The estimates indicate that the gains from coaching are mainly observed at the coaching day and during the days immediately following the coaching session. The

⁷The bus models used in the sample area are reported in Appendix section 3.A.2.

⁸The interaction term between bus type and day fixed effects (ζ_{bt}) is added to the model to account for potential heterogeneity over time in the ease or difficulty of avoiding ABC events due to bus types having different thresholds.

⁹The full regression results for fuel economy are reported in Table 3.A.6.

Figure 3.2: Temporal Effects In-Person Coaching

Notes: Driving performance in the 10 weeks before and after coaching. The day of coaching itself is point 0 on the x -axis. The vertical spikes indicate 95% confidence intervals. The dependent variable fuel economy is measured in liters/100km and acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, number of passengers and bus stops, dummies for routes, days, fill-in rides, additional coaching sessions, bus types, morning and evening rush hours, and the interaction of bus type and day fixed effects. Coaches themselves are excluded from the sample.

overall effect on fuel economy in the post-coaching period, a reduction of 0.08 liters per 100 km ($0.06\sigma_r$), is only marginally significant and not robust when evaluated against MHT-adjusted p -values.

Table 3.4: In-Person Coaching Effects on Driving Performance

Outcome		Coefficient	p -values		
			Unadj.	Bonf.	Holm
Fuel economy	day of coaching	-0.5611	0.0000***	0.0000***	0.0000***
	day 1-7 after coaching	-0.2979	0.0002***	0.0024***	0.0018***
	post-coaching	-0.0814	0.0757*	0.9084	0.8327
Acceleration	day of coaching	-1.8728	0.0000***	0.0000***	0.0000***
	day 1-7 after coaching	-1.1025	0.0000***	0.0000***	0.0000***
	post-coaching	-0.9033	0.0000***	0.0000***	0.0000***
Braking	day of coaching	-0.4247	0.0000***	0.0000***	0.0000***
	day 1-7 after coaching	-0.2947	0.0000***	0.0000***	0.0000***
	post-coaching	-0.3032	0.0000***	0.0000***	0.0000***
Cornering	day of coaching	-0.2533	0.0000***	0.0000***	0.0000***
	day 1-7 after coaching	-0.1300	0.0050***	0.0600*	0.0500*
	post-coaching	0.0249	0.6093	1.0000	1.0000

Notes: Identification of in-person coaching effects on driving performance. The dependent variable fuel economy is measured in liters/100km; acceleration, braking and cornering as the number of events per 10 kilometers. Reported are the DID effects of coaching on the day of coaching, in the first week following coaching and in the entire post-coaching period following coaching. Standard p -values as well as p -values that use a Bonferroni and Holm correction for multiple hypothesis testing are reported. Standard errors are clustered by driver. Full regression results are reported in Tables 3.A.6 and Table 3.A.7 in the appendix.

***(**, *) : the corresponding p -values are less than 1% (5% or 10%).

The improvement in driving behavior is also observed when looking at the ABC driving dimensions, although a word of caution in interpreting the results is that coached drivers already perform slightly better on acceleration and braking prior to coaching. Despite this, the coefficient plots in Figure 3.2 provide suggestive evidence that drivers improve their performance on the ABC dimensions once they have interacted with a coach. Taking the pre-coaching period as the baseline again, the estimates in Table 3.4 show that performance improves on all ABC dimensions at the day of coaching: the decrease in events per 10 kilometers for acceleration, braking, and cornering is 1.87 ($0.80\sigma_r$), 0.42 ($0.38\sigma_r$), and 0.25 ($0.20\sigma_r$),

respectively.¹⁰ Only for acceleration and braking is a significant effect observed in the entire post-coaching period, with a decrease of respectively 0.90 events ($0.40\sigma_r$) and 0.30 events ($0.27\sigma_r$) per 10 kilometers. In a robustness check where I allow for group-specific linear time trends, I find that coaching effects continue to be observed on all four driving dimensions and the effects all remain in the expected direction.¹¹

One concern might be that the previous findings are not driven by inherent features of in-person coaching, but that they are instead the result of idiosyncratic coach characteristics that are hard to copy. In the latter case, this may imply that the treatment effects are only observed for a select few coaches. A direct comparison of differences in coaching is not possible because the coach logs lack this information. I will therefore take the approach used in Romensen and Soetevent (2020a) and estimate in Figure 3.A.3 the treatment effects on fuel economy at the coach level.¹² Even though there is loss of precision in the estimates due to smaller sample sizes, the pattern of temporary improvements in driving behavior is seen among multiple coaches. This corroborates that the improvements in driving are due to in-person coaching features that transcend coach-specific qualities.

Next, I consider whether there is treatment heterogeneity with respect to coaching depending on a driver's past performance. Romensen and Soetevent (2020a) find that the improvements in productivity due to coaching are mainly observed among drivers in the bottom half of the performance distribution. They took the following approach to arrive at this result. First, the monthly relative rank of a driver on each of the four driving dimensions is determined by comparing the driver with his or her peers in the same base location. Then, for every driver coached in month m , the relative performance in the month before coaching ($m - 1$) is compared to the relative performance in the month after coaching ($m + 1$). This comparison is done for all four productivity measures. The change in performance is evaluated against the benchmark of the change in perfor-

¹⁰The full regression results for acceleration are reported in Table 3.A.6. The results for braking and cornering can be found in Table 3.A.7.

¹¹The regression results with group-specific linear time trends can be found in Table 3.A.8 (fuel economy and acceleration) and in Table 3.A.9 (braking and cornering).

¹²The results for the ABC dimensions are shown in Figures 3.A.4-3.A.6.

mance from month $m - 1$ to $m + 1$ that non-coached drivers experience. I follow the same procedure here and first construct the relative rankings of all drivers in the two concession areas that are considered in this chapter. Figure 3.A.7 summarizes the results and shows for coached drivers and non-coached drivers the change in ranking.¹³ The figure distinguishes between drivers that initially (in month $m - 1$) belong to either the top 25%, bottom 50%, or the remaining group in terms of driving performance. I find that drivers that initially are in the bottom 50% are most responsive to the coach and improve the most in terms of their relative ranking. The effects are, however, smaller in magnitude and not significantly different from zero in the case of fuel economy.

Are the effects of in-person coaching different when drivers already received a feedback report in the past? The information that is provided in the report might overlap with what is covered during a coaching session, and this may in turn reduce the effect of coaching on driving behavior. The coach logs indicate that 379 (216) drivers received coaching prior to (after) the delivery of the first report around 15 January 2017. If the previous reasoning holds, then the observed treatment effects of coaching are mainly driven by the first group of coached drivers. In Romensen and Soetevent (2020a), however, it was found that coaching is effective regardless of having been exposed to feedback reports in the past. Importantly, though, in their study the feedback reports contained targeted peer-comparison messages on the ABC driving dimensions. These messages are not part of the reports that are distributed in the sample area of this study. Figure 3.A.8 plots the effects of coaching on fuel economy and acceleration in the period before and after the distribution of the first feedback report. As in Romensen and Soetevent (2020a), in both periods drivers improve on the day of coaching and in the subsequent weeks. Hence, having already been informed on one's own performance through a feedback report does not seem to interfere with the effect of being coached by an experienced peer.

To summarize, the results presented in this study are very much in line with those reported in Romensen and Soetevent (2020a). By evaluating the same coaching program in another sample area and period, I find that bus

¹³The corresponding regression estimates can be found in Table 3.A.10.

drivers immediately improve on their driving at the day of coaching, and this effect persists, albeit at a smaller magnitude, in the weeks following coaching. The effect of coaching is subject to decay and after about six weeks no remaining effect of the coaching session can be detected in the data for fuel economy. For acceleration and braking, there is some evidence of persistence in the positive effects. Drivers in the bottom half of the performance distribution are most responsive to coaching and show the largest improvements in driving behavior. A parallel feedback program in which drivers receive a monthly written feedback report does not attenuate the positive impact of the coaching program.

3.6 Conclusion

Interactions between workers in the workplace may facilitate exchanges of knowledge and skills. Such interactions can help workers to improve on work-related outcomes. Previous research has shown that workers become more productive when they work alongside productive coworkers, though it is less clear if this effect persists if the social pressure mechanism is absent. Furthermore, few studies have examined how firms can design workplace programs that proactively facilitate productive worker interactions. How can workers be matched with high-achieving peers such that peer learning is promoted and that sustained outcome improvements are achieved?

In this chapter, I evaluate the impact of a coaching program on worker productivity at a large public transport company in the Netherlands. The program involves on-the-road coaching sessions to bus drivers by experienced peers. By collecting detailed logs from all coaches operating in two concession areas of the company, I am able to pin down for every driver in the sample when and how often the driver is coached. In a quasi-experimental framework with close to two million trip-level observations from more than 1,000 bus drivers, the impact of a single coaching session is evaluated on four different outcome measures of driving behavior (fuel economy, acceleration, braking, and cornering).

The main results of the evaluation are as follows. First, drivers significantly improve their driving behavior after they participated in a coaching

session. On the day of coaching, the fuel need for driving 100 kilometers decreases with about 0.6 liters per driver. Improved fuel economy is also observed in the six weeks following coaching, but after that no remaining effect can be detected in the data. This suggests that drivers do not habituate to the more efficient driving style. For acceleration and braking, the positive effect of coaching appears more persistent. Second, the improvements due to coaching are mainly observed among drivers that have relatively poor baseline scores on the driving dimensions. Finally, coaching improves driving performance irrespective of whether drivers received performance feedback reports in the past. These results are qualitatively in line with the main findings of a companion study that is reported in Romensen and Soetevent (2020a).

In-person coaching by experienced peers thus seems a promising means to facilitate knowledge and skill transfers between workers on the work floor. Especially workers with low initial productivity stand to benefit the most from receiving tailored feedback from high-achieving coworkers. The observed decay in the treatment effects suggests that more research is needed that investigates how the improvements induced by peer coaching can be made more permanent. Studies that examine which particular elements of in-person coaching lead to the improvements in productivity and sustainable behavior are also welcome. How to shape workplace interactions to foster positive changes in worker performance continues to be a promising point of departure for future research.

3.A Appendix

3.A.1 Sample area

Figure 3.A.1:
Location of the Sample Area in the Netherlands

Notes: The sample area used in this study consists of two bus concessions: Zuid-Holland Noord (area A) and Drechtsteden, Alblasserwaard and Vijfheerenlanden (area B).

3.A.2 Construction final sample

The initial sample consists of 2,846,818 trip-level observations. Table 3.A.1 details the applied cleaning steps. Because of data similarities and to facilitate the comparison of results, the cleaning procedure is similar to the one used in Romensen and Soetevent (2020a).

Table 3.A.1: Cleaning Steps for Sample Construction

		% share of full sample
Full sample	2,846,818	
<i>Reason for dropping observation:</i>		
Duplicate observation (in terms of all variables)	(15,273)	0.54
No bus identifier	(622,092)	21.85
Bus type not eligible for EOBR	n.a.	
Error message from EOBR	(54,690)	1.92
Within-driver obs. with the same departure date/time	(86,416)	3.04
Very short rides (less than 1 kilometer)	(57,313)	2.01
Punctuality: more than 1 hour	(1,027)	0.04
<i>Unreasonable outcomes of dependent variables:</i>		
- Fuel economy: less than 12.5 or more than 100	(27,165)	0.95
- ABC dimensions: more than 5 SDs above the mean	(13,434)	0.47
	1,969,408	

Notes: Fuel economy is measured in liters/100km. The ABC comfort dimensions are the number of events per 10 kilometers (a reduction in events indicates better driving behavior). Trip punctuality is the difference between actual and planned driving time.

The company operates various bus types in the concession areas of this study. Table 3.A.2 provides an overview of the different bus types. Most trips are completed with VDL buses (54.30%). Fuel economy data are available for all bus types because all buses run on diesel. All bus types have the same start dates for the recording of data on driving performance. For fuel economy, the start date is 1 July 2015, and for the ABC dimensions it is 1 September 2016.

Table 3.A.2: Fleet Information

Bus type	length (m)	fuel type	No. trips	% share of full sample
DAF SB200 AMBASSADOR	12.0	diesel	271,071	13.76
VDL AMBASSADOR ALE 106	10.6	diesel	76,923	3.91
VDL CITEA LLE 120	12.0	diesel	992,486	50.40
VOLVO 8900 LE 13M	13.0	diesel	454,966	23.10
VOLVO 8900 LE 14.7M	14.7	diesel	27,405	1.39
VOLVO 7700 B5 LH Hybrid	12.0	diesel	146,557	7.44
Analysis set			1,969,408	
DAF AMBASSADOR			271,071	13.76
VDL			1,069,409	54.30
VOLVO			628,928	31.93

3.A.3 Timing of coaching sessions

Figure 3.A.2: Time of First Coaching

Notes: moment of first coaching for drivers. Dark blue bars indicate the drivers who received their first coaching during a specific week as a share of the total number of drivers operating during that week. The light blue bars depict the cumulative share of coached drivers operating during a week. Announcement refers to the kickoff event on 1 December 2016 during which the coaching program and the electronic on-board recorder (EOBR) were officially announced to the bus drivers.

3.A.4 Determinants of ABC outcomes

Table 3.A.3: Determinants of Acceleration

Dependent variable:	Acceleration				
	(1)	(2)	(3)	(4)	(5)
DAF	-5.889*** (0.289)	-5.869*** (0.286)	-6.245*** (0.290)	-6.059*** (0.250)	-6.207*** (0.269)
VDL 10m	0.793** (0.322)	0.775** (0.321)	1.188*** (0.295)	1.045*** (0.289)	1.131*** (0.175)
Volvo 13m	-6.093*** (0.196)	-6.105*** (0.195)	-6.078*** (0.195)	-6.102*** (0.171)	-6.170*** (0.183)
Volvo 14m	-8.719*** (0.222)	-8.718*** (0.220)	-8.838*** (0.218)	-8.834*** (0.189)	-8.702*** (0.203)
Volvo hybrid	-12.632*** (0.371)	-12.637*** (0.369)	-13.140*** (0.372)	-13.179*** (0.345)	-13.353*** (0.357)
Rush hour 7-10am			-0.124*** (0.046)	-0.130*** (0.045)	-0.154*** (0.025)
Rush hour 4-7pm			0.656*** (0.048)	0.687*** (0.048)	0.774*** (0.030)
Non-scheduled trip			-0.478 (0.311)	-0.174 (0.322)	0.281* (0.153)
No. of stops per km.			0.542*** (0.128)	0.521*** (0.104)	0.517*** (0.089)
Trip length (in km.)			-0.058*** (0.003)	-0.056*** (0.003)	-0.052*** (0.003)
Ln(No. of passengers)			1.566*** (0.034)	1.436*** (0.032)	1.369*** (0.029)
Punctuality			-0.074*** (0.008)	-0.068*** (0.008)	0.025*** (0.003)
Constant	13.200*** (0.164)	13.397*** (0.168)	9.387*** (0.282)	7.609*** (0.365)	12.942*** (1.119)
R ²	.577	.579	.601	.63	.71
Number of trip-level observations	1098013	1098013	1098013	1098013	1098013
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of acceleration events per 10 km. Default categories: trips outside rush hours completed with a VDL 12m bus. Trip punctuality is the difference in minutes between actual and scheduled driving time. Standard errors are clustered by driver.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 3.A.4: Determinants of Braking

Dependent variable:	Braking				
	(1)	(2)	(3)	(4)	(5)
DAF	3.476*** (0.189)	3.505*** (0.188)	3.478*** (0.185)	3.546*** (0.177)	3.367*** (0.174)
VDL 10m	-0.341*** (0.049)	-0.349*** (0.049)	-0.280*** (0.052)	-0.324*** (0.053)	-0.336*** (0.051)
Volvo 13m	2.897*** (0.116)	2.889*** (0.116)	2.917*** (0.114)	2.891*** (0.108)	2.891*** (0.107)
Volvo 14m	3.116*** (0.136)	3.122*** (0.135)	3.133*** (0.131)	3.117*** (0.125)	3.132*** (0.120)
Volvo hybrid	1.719*** (0.207)	1.735*** (0.206)	1.704*** (0.203)	1.680*** (0.196)	1.472*** (0.192)
Rush hour 7-10am			-0.05*** (0.018)	-0.049*** (0.018)	-0.041*** (0.011)
Rush hour 4-7pm			0.189*** (0.024)	0.195*** (0.024)	0.195*** (0.012)
Non-scheduled trip			-0.101 (0.092)	0.014 (0.095)	0.140** (0.065)
No. of stops per km.			-0.090* (0.048)	-0.244*** (0.048)	-0.231*** (0.040)
Trip length (in km.)			-0.012*** (0.002)	-0.013*** (0.001)	-0.012*** (0.001)
Ln(No. of passengers)			0.232*** (0.014)	0.195*** (0.014)	0.174*** (0.011)
Punctuality			-0.036*** (0.003)	-0.033*** (0.003)	0.002 (0.001)
Constant	0.533*** (0.076)	0.673*** (0.076)	0.347*** (0.110)	1.517*** (0.294)	1.830*** (0.542)
R ²	.288	.293	.3	.34	.485
Number of trip-level observations	1098013	1098013	1098013	1098013	1098013
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of braking events per 10 km. Default categories: trips outside rush hours completed with a VDL 12m bus. Trip punctuality is the difference in minutes between actual and scheduled driving time. Standard errors are clustered by driver.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 3.A.5: Determinants of Cornering

Dependent variable:	Cornering				
	(1)	(2)	(3)	(4)	(5)
DAF	0.162*	0.158*	0.177**	0.198**	0.111
	(0.087)	(0.087)	(0.088)	(0.087)	(0.089)
VDL 10m	-0.224**	-0.225**	-0.193**	-0.184**	-0.243***
	(0.087)	(0.088)	(0.086)	(0.086)	(0.053)
Volvo 13m	0.003	0.006	0.032	0.045	0.066
	(0.063)	(0.062)	(0.060)	(0.060)	(0.052)
Volvo 14m	-0.295***	-0.291***	-0.258***	-0.224***	-0.183***
	(0.069)	(0.069)	(0.066)	(0.065)	(0.058)
Volvo hybrid	6.792***	6.784***	6.838***	6.873***	6.798***
	(0.225)	(0.225)	(0.224)	(0.224)	(0.227)
Rush hour 7-10am			-0.135***	-0.134***	-0.127***
			(0.021)	(0.021)	(0.010)
Rush hour 4-7pm			0.113***	0.111***	0.139***
			(0.025)	(0.025)	(0.012)
Non-scheduled trip			-0.179*	-0.178*	-0.105
			(0.099)	(0.097)	(0.085)
No. of stops per km.			0.251***	0.162***	0.175***
			(0.040)	(0.043)	(0.039)
Trip length (in km.)			-0.006***	-0.008***	-0.006***
			(0.001)	(0.001)	(0.001)
Ln(No. of passengers)			0.077***	0.090***	0.067***
			(0.016)	(0.017)	(0.011)
Punctuality			-0.079***	-0.080***	-0.030***
			(0.003)	(0.003)	(0.001)
Constant	1.148***	1.175***	0.470***	2.094***	1.080***
	(0.049)	(0.051)	(0.095)	(0.254)	(0.413)
R ²	.422	.423	.429	.436	.578
Number of trip-level observations	1098013	1098013	1098013	1098013	1098013
Weather dummies	No	Yes	Yes	No	No
Driver fixed effects	No	No	No	No	Yes
Day fixed effects	No	No	No	Yes	Yes
Route fixed effects	No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of cornering events per 10 km. Default categories: trips outside rush hours completed with a VDL 12m bus. Trip punctuality is the difference in minutes between actual and scheduled driving time. Standard errors are clustered by driver.

***(*,*) : statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

3.A.5 In-person coaching effects

Table 3.A.6: In-Person Coaching Effects on Driving Performance: Fuel Economy and Acceleration

Dependent variable:	Fuel Economy			Acceleration		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-coaching	0.196*** (0.061)	-0.081* (0.046)		-0.316*** (0.104)	-0.903*** (0.101)	
Day of first coaching			-0.561*** (0.095)			-1.873*** (0.169)
1 –7 days after			-0.298*** (0.080)			-1.102*** (0.123)
8 –14 days after			-0.295*** (0.081)			-1.169*** (0.131)
15 –21 days after			-0.279*** (0.086)			-0.890*** (0.133)
22 –28 days after			-0.299*** (0.071)			-0.976*** (0.132)
29 –35 days after			-0.275*** (0.079)			-0.850*** (0.130)
36 –42 days after			-0.174** (0.083)			-0.866*** (0.130)
43 –49 days after			-0.101 (0.099)			-0.888*** (0.142)
50 –56 days after			-0.031 (0.102)			-0.864*** (0.145)
57 –63 days after			0.042 (0.098)			-0.858*** (0.149)
64 –70 days after			-0.111 (0.081)			-0.801*** (0.139)
> 70			-0.009 (0.051)			-0.816*** (0.118)
Number of trip-level observations	1788720	1788720	1788720	1100419	1100419	1100419
Controls	No	Yes	Yes	No	Yes	Yes
Driver fixed effects	No	Yes	Yes	No	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	Yes	Yes	No	Yes	Yes

Notes: Identification of in-person coaching effects on driving performance. Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. Fuel economy is defined as liters/100km. Acceleration is measured as the number of events per 10 kilometers. Fewer events means improved driving behavior. Standard errors are clustered by driver. Controls include: travel distance, route dummies, fill-in rides, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, and having been coached before.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 3.A.7: In-Person Coaching Effects on Driving Performance: Braking and Cornering

Dependent variable:	Braking			Cornering		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-coaching	0.237*** (0.057)	-0.303*** (0.038)		-0.114** (0.051)	0.025 (0.049)	
Day of first coaching			-0.425*** (0.057)			-0.253*** (0.050)
1 –7 days after			-0.295*** (0.044)			-0.130*** (0.046)
8 –14 days after			-0.328*** (0.044)			-0.133** (0.052)
15 –21 days after			-0.326*** (0.048)			-0.068 (0.057)
22 –28 days after			-0.277*** (0.048)			-0.007 (0.051)
29 –35 days after			-0.342*** (0.046)			-0.031 (0.058)
36 –42 days after			-0.287*** (0.050)			-0.025 (0.056)
43 –49 days after			-0.274*** (0.055)			-0.024 (0.055)
50 –56 days after			-0.298*** (0.049)			0.028 (0.058)
57 –63 days after			-0.302*** (0.054)			0.029 (0.068)
64 –70 days after			-0.291*** (0.048)			0.075 (0.058)
> 70			-0.298*** (0.044)			0.100* (0.059)
Number of trip-level observations	1100419	1100419	1100419	1100419	1100419	1100419
Controls	No	Yes	Yes	No	Yes	Yes
Driver fixed effects	No	Yes	Yes	No	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	Yes	Yes	No	Yes	Yes

Notes: Identification of in-person coaching effects on driving performance. Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. The dependent variables braking and cornering are measured as the number of events per 10 kilometers. Fewer events means improved driving behavior. Standard errors are clustered by driver. Controls include: travel distance, route dummies, fill-in rides, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, and having been coached before.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 3.A.8: In-Person Coaching Effects on Driving Performance: Fuel Economy and Acceleration (Controlling for Linear Trend)

Dependent variable:	Fuel Economy			Acceleration		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-coaching	0.085 (0.063)	-0.126** (0.051)		-0.569*** (0.109)	-0.805*** (0.100)	
Day of first coaching			-0.556*** (0.094)			-1.792*** (0.167)
1 –7 days after			-0.293*** (0.080)			-1.015*** (0.120)
8 –14 days after			-0.290*** (0.082)			-1.065*** (0.129)
15 –21 days after			-0.273*** (0.087)			-0.777*** (0.133)
22 –28 days after			-0.293*** (0.073)			-0.856*** (0.130)
29 –35 days after			-0.269*** (0.081)			-0.722*** (0.128)
36 –42 days after			-0.167** (0.082)			-0.728*** (0.130)
43 –49 days after			-0.094 (0.101)			-0.737*** (0.141)
50 –56 days after			-0.023 (0.104)			-0.694*** (0.146)
57 –63 days after			0.049 (0.100)			-0.682*** (0.150)
64 –70 days after			-0.103 (0.084)			-0.617*** (0.143)
> 70			0.002 (0.063)			-0.501*** (0.137)
Number of trip-level observations	1788720	1788720	1788720	1100419	1100419	1100419
Controls	No	Yes	Yes	No	Yes	Yes
Driver fixed effects	No	Yes	Yes	No	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	Yes	Yes	No	Yes	Yes
Linear trend	Yes	Yes	Yes	Yes	Yes	Yes
Coached × linear trend	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Identification of in-person coaching effects on driving performance. Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. Coached is an indicator for drivers who participated in a coaching session during the sample period. Fuel economy is defined as liters/100km. Acceleration is measured as the number of events per 10 kilometers. Fewer events means improved driving behavior. Standard errors are clustered by driver. Controls include: travel distance, route dummies, fill-in rides, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, and having been coached before.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 3.A.9: In-Person Coaching Effects on Driving Performance: Braking and Cornering (Controlling for Linear Trend)

Dependent variable:	Braking			Cornering		
	(1)	(2)	(3)	(4)	(5)	(6)
Post-coaching	0.064 (0.046)	-0.247*** (0.037)		-0.151*** (0.043)	-0.080* (0.041)	
Day of first coaching			-0.394*** (0.057)			-0.280*** (0.049)
1 –7 days after			-0.261*** (0.044)			-0.159*** (0.044)
8 –14 days after			-0.289*** (0.044)			-0.168*** (0.049)
15 –21 days after			-0.283*** (0.047)			-0.105* (0.054)
22 –28 days after			-0.231*** (0.047)			-0.047 (0.047)
29 –35 days after			-0.294*** (0.047)			-0.073 (0.055)
36 –42 days after			-0.235*** (0.051)			-0.070 (0.053)
43 –49 days after			-0.217*** (0.056)			-0.074 (0.054)
50 –56 days after			-0.234*** (0.049)			-0.028 (0.056)
57 –63 days after			-0.236*** (0.054)			-0.029 (0.067)
64 –70 days after			-0.222*** (0.048)			0.014 (0.058)
> 70			-0.178*** (0.046)			-0.004 (0.053)
Number of trip-level observations	1100419	1100419	1100419	1100419	1100419	1100419
Controls	No	Yes	Yes	No	Yes	Yes
Driver fixed effects	No	Yes	Yes	No	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	Yes	Yes	No	Yes	Yes
Linear trend	Yes	Yes	Yes	Yes	Yes	Yes
Coached × linear trend	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Identification of in-person coaching effects on driving performance. Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. Coached is an indicator for drivers who participated in a coaching session during the sample period. The dependent variables braking and cornering are measured as the number of events per 10 kilometers. Fewer events means improved driving behavior. Standard errors are clustered by driver. Controls include: travel distance, route dummies, fill-in rides, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, and having been coached before.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

3.A.6 Treatment effects of coaching at the coach level

Figure 3.A.3: Temporal Effects of In-Person Coaching at Coach Level: Fuel Economy

Notes: Coaches #2 and #3 are stationed at the same base location (the largest one in the sample area). The coach logs are unfortunately not detailed enough to specify per driver which of the two coaches was involved in the coaching session. The figure therefore shows the combined effect for the two coaches.

Figure 3.A.4: Temporal Effects of In-Person Coaching at Coach Level: Acceleration

Notes: Coaches #2 and #3 are stationed at the same base location (the largest one in the sample area). The coach logs are unfortunately not detailed enough to specify per driver which of the two coaches was involved in the coaching session. The figure therefore shows the combined effect for the two coaches.

Figure 3.A.5: Temporal Effects of In-Person Coaching at Coach Level: Braking

Notes: Coaches #2 and #3 are stationed at the same base location (the largest one in the sample area). The coach logs are unfortunately not detailed enough to specify per driver which of the two coaches was involved in the coaching session. The figure therefore shows the combined effect for the two coaches.

Figure 3.A.6: Temporal Effects of In-Person Coaching at Coach Level: Cornering

Notes: Coaches #2 and #3 are stationed at the same base location (the largest one in the sample area). The coach logs are unfortunately not detailed enough to specify per driver which of the two coaches was involved in the coaching session. The figure therefore shows the combined effect for the two coaches.

3.A.7 Treatment heterogeneity of coaching

Figure 3.A.7: Differential Treatment Effects of Coaching

Note: Green (blue) : Average ranking of drivers (not) coached in month m in month $m-1$ (x -axis) and month $m+1$ (y -axis). The lines plot the non-parametric piece-wise local polynomial fit using an Epanechnikov kernel function. For non-coached drivers, the 95% confidence interval (shaded area) is shown as well.

Table 3.A.10: Change in ranking month $(m + 1)$ vs. $(m - 1)$ following coaching in month m

Dep. var	Rank month $(m + 1)$			
		Acceleration		
Coached month m		0.020 (0.022)	-0.013 (0.024)	-0.032* (0.019)
Rank month $(m - 1)$		0.958*** (0.083)	0.733*** (0.062)	0.627*** (0.033)
obs.		1,272	1,937	3,420
R^2		0.1529	0.0785	0.1733
Sample	Rank month $(m - 1)$:	top-25%	25-50%	bottom-50%
		Braking		
Coached month m		-0.022 (0.017)	-0.002 (0.030)	-0.026* (0.016)
Rank month $(m - 1)$		0.951*** (0.078)	0.750*** (0.063)	0.684*** (0.030)
obs.		1,408	1,879	3,342
R^2		0.1409	0.0857	0.2114
Sample	Rank month $(m-1)$:	top-25%	25-50%	bottom-50%
		Cornering		
Coached month m		0.007 (0.014)	-0.001 (0.020)	-0.037** (0.016)
Rank month $(m - 1)$		0.885*** (0.055)	0.862*** (0.057)	0.901*** (0.022)
obs.		1,516	1,799	3,314
R^2		0.2031	0.1481	0.4396
Sample	Rank month $(m - 1)$:	top-25%	25-50%	bottom-50%
		Fuel economy		
Coached month m		-0.013 (0.024)	-0.023 (0.025)	-0.018 (0.017)
Rank month $(m - 1)$		0.796*** (0.098)	0.565*** (0.063)	0.713*** (0.035)
obs.		1,354	2,000	3,275
R^2		0.0883	0.0435	0.2057
Sample	Rank month $(m - 1)$:	top-25%	25-50%	bottom-50%

Notes: All regressions include a constant. Standard errors clustered by driver.

***(*, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

3.A.8 Treatment effects of coaching with/without prior feed-back exposure

Figure 3.A.8: Impact In-Person Coaching With and Without Having First Received Written Feedback Reports

Chapter 4

Rabbits and study habits:

A field experiment on pacesetters and student effort*

Abstract

Many students perform below their academic potential due to a lack of study time. Targeting study time and evaluating the effects has thus far been challenging because of poor measurement of effort and the absence of information on students' initial study plans. In a field experiment with 573 university students, I elicit detailed weekly study goals and plans from students and use these to construct individualized pacesetters (rabbits). Pacesetters are moving reference points that visualize the preferred study pace of the present self by moving exactly according to the initial study plan. I build a new educational technology to measure effort and to display the pacesetters to students in real time. Falling behind the pacesetter confronts the student with his or her procrastination. I find that students have more ambitious study goals when they set a pacesetter, but subsequently tend to fall behind quickly and are less likely to reach their goals. The pacesetter has no impact on study time, nor does it improve learning outcomes. I discuss treatment heterogeneity based on gender, ability, and procrastination.

*This chapter is based on Romensen (2020b). The educational technology that is used in this chapter is co-developed with Tonny Romensen.

4.1 Introduction

A student's own study effort is a core input in the education production function and is the one input that is under direct control of the student. Various studies stress and show that a lack of study effort has detrimental effects on educational achievements (Metcalf, Burgess, and Proud 2019, Stinebrickner and Stinebrickner 2008, Angrist, Lang, and Oreopoulos 2009). Observed declines in students' study effort are therefore especially worrisome (Oreopoulos, Patterson, Petronijevic, and Pope 2018). Between 1961 and 2004, the average weekly study time of college students has decreased by about a half (Babcock and Marks 2011). In light of OECD figures showing that delayed graduation is common and that about 31% of the students drop out of tertiary education (OECD 2008), the problem of study effort decline thus seems acute and in urgent need of attention from researchers.

How can students be motivated to study more? Policies that can raise student effort may have great promise in improving various educational outcomes (Lavecchia, Liu, and Oreopoulos 2016). Implementing such policies, however, is widely considered a challenge. It requires detailed data on effort and students' initial study plans in order to provide timely feedback on individual progress. Such data are often absent in existing research. Furthermore, students tend to have self-control problems in how effort is allocated over time (Augenblick, Niederle, and Sprenger 2015, Wong 2008). Procrastination is widespread and estimated to affect up to 95% of tertiary students (Steel 2007). These issues have stalled progress on designing and implementing policies that can meaningfully alter effort.

In this chapter, I aim to overcome these issues by developing and leveraging a new educational technology, called the *MicroApp* (an online learning platform), to meticulously measure student effort and to administer personalized planning interventions to students in real time. Specifically, in a field experiment I examine the effect on student effort when prior study commitments by students are presented as real-time virtual pacesetters (*rabbits*). In sports, a rabbit is the colloquial term used for a runner who is instructed to run at the exact pace needed to finish at a certain time. In education, pacesetters can provide a succinct and salient reference point

that visualizes the preferred study pace envisioned by the present self. The distant study goal is transformed into a continuous and easy-to-interpret task that activates students directly: stay close to your own pacesetter if you want to reach your initial study goal.

As the pacesetter is, by design, not prone to self-control problems, falling behind immediately makes salient to students that they are renegeing on earlier commitments. Under the assumption that students exhibit reference-dependent preferences, loss aversion makes falling behind psychologically painful and motivates the student to catch up and exert effort. The pacesetter serves as a commitment whereby future selves are induced to follow the pacesetter and, in doing so, implement the initial study plan.

I use the phase-in of the MicroApp at a Dutch university to implement a field experiment that identifies the causal effects of the pacesetter on effort and learning outcomes. The experiment covers a large introductory course in microeconomics with a total enrollment of 573 students. The MicroApp is a self-developed online learning platform that is specifically created for this course. The platform contains course recaps and ample practice material with feedback. Students can use the platform during self-study and select at any time the course topics they want to work on. The MicroApp is offered free-of-charge to students and is a non-mandatory component of the course. This way, I can treat platform usage as a measure of pure study time, and not as a measure of actions undertaken to directly satisfy some course criterium. Student activity on the platform is minutely recorded and registered in individual user logs.

In the experiment, students are asked each week through in-class surveys to formulate a study goal and plan regarding MicroApp usage in the upcoming week. I ask students in the *control* group only to formulate a study goal. Their MicroApp version will show how much of that goal is achieved at every moment. I ask students in the *planner* group for a detailed study plan in terms of MicroApp usage: when and how long do they plan to use the platform? Their MicroApp version allows them to check back on the study plan they submitted. Students in the *rabbit* group also submit a study plan, but their MicroApp has the pacesetter feature: it visualizes how much time they spent so far this week, and how much time

they were planning to have spent. In other words, they observe in real time whether they are ahead or behind their own pacesetter.

By comparing usage of the MicroApp across all three conditions, a detailed picture is created of student effort and the distribution of this effort over time. Matching these data with other survey information and administrative databases, I can check how usage affects other learning inputs (e.g., attending lectures) and learning outcomes (e.g., final exam score).

The results are as follows. The pacesetter induces students to set more ambitious study goals, amounting to about two additional study hours per week. Similar effects are observed, albeit smaller, for students in the planner condition. Interestingly, students in both treatments are less likely to complete their study goals on time, partly as a result of having set too ambitious study goals. I find no evidence that the treatments increase study time on the MicroApp, nor do I observe effort spillovers to other learning inputs or other courses. Further analyses reveal no treatment impact on various learning outcomes. Treatment heterogeneity based on gender, ability, and procrastination is considered in detail but found to be only limited in terms of effort, goals, and learning outcomes. The null results are in line with research on low-cost interventions that target study time (Oreopoulos, Patterson, Petronijevic, and Pope 2018).

This study is one of the first attempts to use educational technologies as a means to target behaviors that have a negative impact on academic performance. I show the potential of using these technologies to measure student effort, to test interventions rigorously by implementing field experiments, and to gather insights on how students learn. The digitalization of education sets a new research agenda and unveils many research opportunities. It will require from researchers and educators alike to think clearly about how students learn in an increasingly digital world.

This chapter proceeds as follows. Section 4.2 discusses related literature. The research question and corresponding hypotheses are covered in Section 4.3. In Section 4.4, I elaborate on the experimental design and setting. The empirical evaluation is presented in Section 4.5. Section 4.6 provides a brief further discussion of the results and details the implications. The conclusions can be found in Section 4.7.

4.2 Related literature

Studying is a classic investment decision, as upfront costs need to be incurred first before benefits can be enjoyed. Whether students can properly execute such a decision is subject to much research (Lavecchia, Liu, and Oreopoulos 2016, Koch, Nafziger, and Nielsen 2015). Central to the discussion is evidence that many students display self-control problems with regard to how effort is allocated over time (Augenblick, Niederle, and Sprenger 2015, Wong 2008). They tend to have difficulties following through on earlier commitments and as a result may provide insufficient effort to reach initial goals. Many students perform poorly in educational programs due to a simple lack of preparation (Oreopoulos, Patterson, Petronijevic, and Pope 2018, Angrist, Lang, and Oreopoulos 2009).

Students are at least partially aware of their own self-control problems (Wong 2008), and they may therefore seek and benefit from commitment devices (Bryan, Karlan, and Nelson 2010, O'Donoghue and Rabin 1999). In education, a commonly used commitment device is imposing a deadline on task completion. For example, evenly spread deadlines may alleviate self-control problems by inducing students to start on a task earlier. Evidence for this is mixed, however, with some studies reporting positive effects (Ariely and Wertenbroch 2002) and others reporting no effects (Burger, Charness, and Lynham 2011). This suggests that deadlines still offer too much leeway for intertemporal effort substitution. Other settings provide further suggestive evidence of this: people bunch their effort right before the deadline when filing taxes (Martinez, Meier, and Sprenger 2017) and paying traffic fines (Heffetz, O'Donoghue, and Schneider 2016).

Recent studies propose goal bracketing, an internal commitment device, to address the issue of study effort being postponed. With goal bracketing, individuals endogenously set subgoals and adopt mental accounts to evaluate their performance in a broad or narrow performance account (Koch and Nafziger 2016, Koch and Nafziger 2017, Hsiaw 2015). In theory, narrow goals (e.g., daily instead of weekly study goals) can be particularly effective for procrastinating students because they rule out behavior in which lack of effort today can be compensated for with more effort in the future. The

trade-off is that it comes at the cost of reduced flexibility. Furthermore, there may be implementation issues. Facilitating goal bracketing requires high-frequency data on effort provision at the student level in order to verify whether goals are met. Also, if narrower goals insulate students better from self-control problems, there may be a risk of feedback with excessive goal bracketing. Providing progress updates on too many goals may overwhelm students and dilute the motivational effects of goal-setting.

In most of the existing research on ways of increasing student effort, actual effort by students is unobserved due to data limitations. Output measures (e.g., exam performance) are used instead to approximate the provided effort. This leads to two issues. First, learning outputs do not reveal how student effort was distributed over time, and to what extent this realized distribution matches initial study plans. Second, ability may confound results when learning outputs are used as proxies for effort levels. Disentangling the effects of ability and effort is crucial for the design of educational interventions (Levitt, List, Neckermann, and Sadoff 2016). When the goal is to motivate students, it may be promising to divert attention from output and to directly target and incentivize learning inputs instead (Clark, Gill, Prowse, and Rush 2017, Lavecchia, Liu, and Oreopoulos 2016).

Educational technologies show considerable promise in tailoring interventions that target behaviors with a negative impact on academic performance (Escueta, Nickow, Oreopoulos, and Quan forthcoming). Student effort and performance can be recorded at low cost by these technologies. This makes it possible to target interventions at learning inputs, and to assess the impact with the help of user data. Technologies also facilitate new ways of visualizing feedback, which makes it easier for students to digest greater amounts of information on progress. Furthermore, by asking students beforehand what their study plans are, technologies can precisely map to what extent actual effort over time matches initial study plans. To date, however, there is little research available that has explored how and which behavioral interventions can be integrated in educational technologies in order to examine and address self-control problems among students.

This study aims to fill this gap. I elicit initial study plans from students and use a novel educational technology to measure how actual effort com-

pares with these plans. I then build on existing goal-setting research and propose an individualized self-study pacesetter (a moving reference point) to keep students on track. The pacesetter preserves flexibility (students can catch up), but still triggers the motivating effect of loss aversion when students fall behind. It has the additional benefit of condensing a large amount of information into a single focus point. A field experiment is implemented to identify the treatment effects of the pacesetter.

Conceptually, a useful way of thinking about the pacesetter feature is to see it as a visualization of the preferred study pace envisioned by the present self. A common view in the literature on intertemporal decision making is that there are conflicts between different selves (Hershfield 2011, Frederick 2003), as in models where an individual is both a long-term planner and a short-sighted doer (Benabou and Pycia 2002, Thaler and Shefrin 1981).¹ Such conflicts may be resolved when the different selves can better connect to each other. Technology may provide the means to achieve this. Hershfield et al. (2011), for example, use ageing software and find that people save more for the future when presented age-progressed renderings of themselves. In this chapter, future selves are triggered to connect with the present self through technology-enabled virtual pacesetters.

Clark et al. (2017) introduce a model of task-based goal-setting that addresses the conflict between different selves in the context of study effort. In their model, a student “planner” first sets a goal $g \geq 0$ for the number of study tasks to complete $a \geq 0$. Performance on the exam increases in $a : f'(a) > 0$. The loss-averse student “actor” then chooses the level of a and suffers goal disutility when a falls short of $g : -\lambda \max\{g - a, 0\}$, where λ captures the strength of goal disutility from loss aversion. Finally, the utility of the student “beneficiary” increases in exam performance. It follows that for a student actor exhibiting quasi-hyperbolic discounting (Laibson 1997), the utility function becomes: $\mu_{act}(a|g) = \beta \delta f(a) - [\lambda \max\{g - a, 0\} + C(a)]$, where $C(a)$ is the cost of effort, $\beta \in (0, 1)$ is the present bias, and $\delta \in (0, 1]$ is the standard discount factor. Crucially, unlike performance-based goals,

¹In their discussion of long-term educational decisions, Lavecchia et al. (2016) also propose a dual-system approach, with one forward-looking system and one that is not. Already in 1759, Adam Smith wrote that behavior is determined by the struggle between the “passions” and the “impartial spectator” (Ashraf, Camerer, and Loewenstein 2005).

the utility function shows that task-based goals result in immediate disutility for the student actor when $a < g$. Temporal distance ($\beta\delta$) that would dampen the motivational effect of the goal is thereby eliminated.² The pacesetter feature does exactly that and shows how temporal distance can be removed by letting students “race” against their own effort-based goals.

4.3 Research question and hypotheses

This chapter aims to address the following question: do students exert higher study effort when they can observe in real time how they perform relative to their own preformulated study goals and plans? I operationalize this research question by evaluating three hypotheses. These hypotheses relate to student effort, meeting study goals, and learning outcomes.

The first and main hypothesis is that students who observe their own pacesetter will exert higher study effort than students who do not observe their own pacesetter. The pacesetter is a salient moving reference point that visualizes a student’s preferred study pace to reach a prespecified study goal. Falling behind the pacesetter immediately makes salient to the student that he or she is reneging on earlier commitments. Under the assumption of reference-dependent preferences, this will be psychologically painful to the student because of loss aversion, and hence will induce the student to catch up and exert more effort. This real-time feedback mechanism is not available to students without the pacesetter. All else the same, I thus expect these control students to exert lower effort. As will be discussed in more detail in the next sections, I evaluate this hypothesis using data from a new online learning platform (MicroApp) and in-class survey responses on time investments in various learning inputs.

If students are indeed triggered not to fall behind their own pacesetter, the implication is that they will be more successful in implementing their study plan and reaching their initial study goal. The second hypothesis is therefore that students who observe their own pacesetter are more successful in reaching their study goals than students who do not observe their own pacesetter. That is, the former group has higher goal completion rates.

²Clark et al. (2017) show formally that this leads to increased study effort.

I will evaluate the second hypothesis using the MicroApp data and detailed weekly data on study goals and plans, elicited through the in-class surveys.

Arguably, the ultimate aim of increasing student effort in a course is to have students better prepared for the exam. If the pacesetter is successful in increasing effort among students (first hypothesis), then I would expect this to have a positive impact on learning outcomes as well. The third hypothesis therefore reads as follows: students who observe their own pacesetter will exert higher study effort and as a result perform better in course examinations than students who do not observe their own pacesetter. I will match the data on student effort with administrative databases to evaluate the third hypothesis.

4.4 Experimental design

In order to evaluate the three hypotheses in the previous section and to answer the research question, I will implement a field experiment in a large course at a Dutch university. In this section, I elaborate on the design of the experiment. I start by discussing the institutional setting and the phase-in of the MicroApp. I then discuss the experimental conditions and the randomization of students into these conditions. Finally, I cover the implementation plan and timeline, as well as the data collection procedures.

4.4.1 Institutional setting

The experiment is implemented during the 2018-2019 academic year in a large first-year introductory course in microeconomics at the University of Groningen, the Netherlands. The course runs during the first block of each academic year. The first block consists of an introduction week, seven lecture weeks, and an exam week. In total, 573 students are enrolled in the course. The majority of these students are regular first-year students (about 76%). The other students are either minor finance students (12%) or pre-MSc students (12%), for whom the course is a mandatory elective.³

³Minor finance students are senior students who come from an educational program that does not give direct access to the MSc Finance program at the Faculty of Economics & Business (FEB), University of Groningen. These students first need to complete a

The course has the following setup. Each lecture week, students participate in a lecture, tutorial, and practical. The lecture is on Tuesday and covers the main course topics of that week. During tutorials on Wednesday or Thursday, a lecturer discusses selected exercises from the textbook. Practicals are on Friday and are led by teaching assistants. Each practical starts with a short multiple-choice quiz on the material of that week. During the second part of the practical, students work on practice exercises.

About halfway through the course, there is a midterm exam that is worth 30% of the final grade. In the exam week, there is a final exam that is worth 60% of the final grade. The remaining 10% of the final grade is the average of the six best practical quizzes (out of seven quizzes in total). In order to receive the remaining 10%, a student needs to participate in at least five (six) out of seven tutorials (practicals).

4.4.2 Phase-in of the MicroApp

The experiment makes use of a novel online learning platform, called the *MicroApp*, that is newly introduced in the course. The MicroApp (plus content) is developed by myself with the aim of offering students an online platform where they can practice and review the course material at their own pace and level. Students have their own password-protected account. They can design and implement individual self-study plans by selecting the course topics they want to work on. The MicroApp has an algorithm that automatically matches these selections with the available practice and recap material. At the time of the study, the MicroApp had 420 practice tasks (60 tasks per lecture week) and 26 recaps that summarize the course topics. Appendix section 4.A.1 provides a few screenshots of the pre-experimental version of the MicroApp.⁴

minor program successfully before they can be enrolled in the MSc Finance program. The minor consists of courses, including the microeconomics course discussed in the text, that will help students overcome the deficiencies. Pre-MSc students are senior students who come from a different faculty or institution. These students first need to successfully complete a tailored Pre-MSc program before they can enter the MSc program of their choice at the FEB. The microeconomics course is often a mandatory component of these Pre-MSc programs.

⁴This version was pre-tested in a similar microeconomics course at another faculty of the University of Groningen. Students gave the platform a perfect score (4 out of 4) in

Following the typology in Escueta et al. (forthcoming), the MicroApp combines elements of online courses, computer-assisted learning, and support for students in dealing with behaviors that negatively impact academic performance. For the purpose of this study, behavioral support will be limited to the interventions of the experiment, which are discussed in Section 4.4.3. An advantage that comes with the inhouse development of the MicroApp is that the interventions can be administered and precisely evaluated through A/B-testing.

The MicroApp will be offered free-of-charge to the students and is not a graded component in the course. This is because I would like to have a measure of student effort as a pure learning input, and not a measure of actions undertaken to directly satisfy some course criterium. Usage of the MicroApp is recorded in real time and saved in individual user logs, allowing me to pin down for every student when, how long, and in what way they made use of the platform. Student identifiers enable matching with administrative databases on background characteristics and test scores.

Students are notified during the introductory lecture of the course that, since the MicroApp is a new and work-in-progress educational technology, different versions may be made available and tested during the course in order to find the “optimal user experience”. It is made clear that everyone will have access to the same course content on the MicroApp and that there are only slight variations in the lay-out and additional features. Students are notified that there will be weekly in-class surveys on time-use during the practicals for “general research purposes” and for improving the course. Finally, it is communicated that it is completely determined by chance which MicroApp version and in-class survey is made available to each student.

What is *not* communicated to the students during the introductory lecture is that the experimental variation is only introduced *after* the midterm exam. That is, all students are initially granted access to the version shown in Appendix section 4.A.1. There are two reasons for a delayed start of the experiment. First, students need some time to become acquainted with the course structure and material (including the MicroApp). Since the microeconomics course is a first-year course offered at the start of the academic

the course evaluation. This indicates that students see the added value of the platform.

year, this reason holds particularly true in the setting of this study. Second, a delayed start allows me to assess MicroApp usage in the absence of interventions. For example, is there bunching of student effort right before the midterm exam (a possible indication of self-control problems)? This first stage of the study yields a convenient baseline measure of student effort.

The in-class survey administered during the practicals before the midterm exam is included in Appendix section 4.A.2. The survey has two questions. The first question is about the microeconomics course in which the MicroApp is implemented. It deals with a student's time investment in various learning inputs, such as attending classes, usage of the MicroApp, reading offline course material, et cetera. The second question concerns a student's time investment in the other courses in which he or she participates. Responses to these questions allow me to examine the relative magnitude of each input and whether substitution between the inputs takes place. Since the survey is administered every week, I can see how time investments evolve throughout the course.

4.4.3 Experimental conditions

In the second stage of the study, I use the MicroApp and the in-class surveys to implement a field experiment in which I explore how students can be best motivated to stick to their initial study plan.⁵ I consider three experimental conditions, which are summarized in Table 4.1. In all three conditions, a *broad goal* regarding MicroApp usage in the upcoming week is elicited through an additional section in the in-class survey on time-use. This broad goal indicates for each student the ex-ante preferred study time on the platform in the upcoming week. In the second and third condition, I additionally ask students in the survey to formulate a detailed *study plan*. In this plan students indicate how they are going to allocate the hours over the week in order to reach the study goal. In the third condition, the study plan is used as input to construct a real-time virtual *pacesetter* on the MicroApp. The pacesetter ("rabbit") is an individualized moving reference point that moves exactly according to the study plan submitted

⁵Using the typology in Harrison and List (2004), the experiment in this study is best described as a natural field experiment.

by a student. Appendix section 4.A.3 shows the experimental variation in the in-class surveys. The relevant survey responses are processed the same day in the MicroApp and are made available to students the next day on.

Table 4.1: Summary of Experimental Conditions

	Features		
	Broad goal	Study plan	Pacesetter
Condition I: control condition	✓		
Condition II: “planner” condition	✓	✓	
Condition III: “rabbit” condition	✓	✓	✓

Notes: *Broad goal* refers to whether students are asked in weekly in-class surveys on time-use to formulate a goal with regard to how many hours they intend to use a new online learning platform (the MicroApp) in the upcoming week. Some students are also asked to formulate a *study plan* in which they make clear how they are going to allocate the hours over the week in order to reach the study goal. The *pacesetter* feature is an individualized moving reference point that moves exactly according to the study plan of the student.

The benefit of this build-on experimental design is that I can examine the marginal impact of formulating a study plan and observing the pacesetter beyond the effect of simply setting a broad study goal for the week (as in the control condition). This greatly facilitates the understanding of what will drive the results. For example, a comparison between the outcomes in the control condition and the planner condition makes it possible to pin down the stand-alone effect of formulating a detailed study plan for the week. In a similar way, I can determine the stand-alone effect of the pacesetter.

In the first experimental condition, the *control* condition, students only have access to a baseline version of the MicroApp. The baseline version contains all recaps and practice material (including feedback), but only general information on progress toward the stated study goal. Students in this condition complete a short version of the in-class survey on time-use, which does not elaborate on the study plan for the upcoming week. Figure 4.1 shows how the control condition is visualized in the MicroApp.

In the second condition, the *planner* condition, students do receive a request in the in-class survey to write down their study plan for the upcoming week with regard to MicroApp usage. Specifically, I ask the students to formulate daily study goals in hours and how they wish to allocate these

hours over the day. The planner version of the MicroApp is the same as the baseline version, but additionally contains information on the study plan. Figure 4.2 shows the planner version of the MicroApp.

The third condition, the *rabbit* condition, is the same as the planner condition with the exception that the study plan is used to construct an individualized virtual pacesetter. The rabbit version of the MicroApp visualizes the pacesetter in real time as a feedback feature. Figure 4.3 shows the rabbit version of the MicroApp.

4.4.4 Randomization and exclusion rules

All students who are enrolled in the course (573 students in total) are randomly allocated to one of the three experimental conditions. The randomization is stratified based on the study phase of the students. As mentioned in Section 4.4.1, most students are regular first-year students, but some are senior students enrolled in a Pre-MSc program or minor finance program. Panel A in Table 4.2 shows that the randomization is successful in balancing the study phase composition in the three experimental conditions.

Table 4.2: Study Phase Composition of Students in the Exp. Conditions

	Study phase			
	All students	Regular (first-year)	Minor finance	Pre-MSc
<i>Panel A: sample based on all enrolled students in the course</i>				
Condition I: control condition	190	145	22	23
Condition II: “planner” condition	191	145	22	24
Condition III: “rabbit” condition	192	145	23	24
Number of students	573	435	67	71
<i>Panel B: sample after imposing exclusion rules</i>				
Condition I: control condition	104	73	15	16
Condition II: “planner” condition	99	68	14	17
Condition III: “rabbit” condition	106	75	14	17
Number of students	309	216	43	50

Notes: All students enrolled in a first-year introductory course in microeconomics at the University of Groningen are randomly allocated to one of the three experimental conditions discussed in Section 4.4.3. The randomization is stratified along the study phase of the enrolled students. Most students are regular first-year students. Some students are senior students who take the course as a mandatory elective in the minor finance program or Pre-MSc program. Please refer to footnote 3 in this chapter for further information on these two programs. The sample used in the empirical analyses is obtained after applying two exclusion rules. First, a student is excluded when he or she participated in less than two practicals after the midterm exam. Second, a student is also excluded when he or she never made use of the MicroApp in the period until the first practical after the midterm exam.

I exclude a student from the empirical analyses when he or she meets one of the following two criteria. These criteria were specified in the pre-analysis

Figure 4.1: MicroApp Version for Students in the Control Condition

(a) Front page shows general progress toward the *study goal* of the student

(b) The *study plan* feature is not available for students in the control condition

Notes: Students in the control condition have access to a baseline version of the MicroApp (a new online learning platform in the course). The baseline version only shows general progress toward the study goal of the student. The study goal is elicited weekly in in-class surveys and is about intended MicroApp usage in the upcoming week. In the control condition, students are not requested to submit a study plan in which they indicate how they are going to allocate study hours over the week in order to reach the study goal. The study plan feature is therefore not available in the MicroApp (“under construction”).

Figure 4.2: MicroApp Version for Students in the Planner Condition

(a) Front page shows general progress toward the *study goal* of the student (b) The *study plan* feature is available for students in the planner condition

Notes: Students in the planner condition have access to the same version of the MicroApp as students in the control condition. The only difference in the MicroApp being that students in the planner condition can request and review their initially submitted study plan at any time. This study plan is elicited weekly in the in-class surveys, together with the corresponding study goal. The study goal is about intended MicroApp usage in the upcoming week and the study plan is about how study hours are allocated over the week in order to reach the study goal.

Figure 4.3: MicroApp Version for Students in the Rabbit Condition

(a) Front page shows general progress toward the *study goal* of the student as well as progress relative to a student's own pacesetter (*rabbit*)

(b) The *study plan* feature is available for students in the rabbit condition

Notes: Students in the rabbit condition complete the same in-class survey questions and have access to the same version of the MicroApp as students in the planner condition. The only difference in the MicroApp is that in the rabbit condition students additionally can track their study progress relative to their own pacesetter (rabbit). The pacesetter is an individualized moving reference point that moves exactly according to the study plan submitted by a student.

plan of this study. First, the student participated in less than two practicals after the midterm exam. The reason for this first criterium is because the interventions are based on input from the in-class surveys administered after the midterm exam. When a student is frequently absent during these classes, the interventions cannot be based on up-to-date information from the students. Second, the student never made use of the MicroApp in the period until the first practical after the midterm exam (when the first in-class survey with experimental variation was administered). With no previous user experience, a student cannot formulate a well-grounded study goal and plan with regard to MicroApp usage. I consider a student having used the MicroApp when he or she solved at least one task on the platform.

Panel B in Table 4.2 shows that balance in terms of study phase composition is preserved when the two exclusion rules are applied. About 83% of the students participated in at least two practicals after the midterm exam, of which 65% were also active users of the MicroApp. All data and results in the remainder of this chapter are based on this restricted sample (unless stated otherwise), as these are the students who actively participate in the course and also make use of the MicroApp. As will be discussed in Section 4.5.1, additional balance tests reveal no significant differences between the experimental groups in terms of a wide range of pre-experimental characteristics, indicating that the randomization is successful.

4.4.5 Implementation plan and timeline of the experiment

Table 4.3 summarizes the timeline of the experiment by listing all notable events. The experiment consists of two stages. In the first stage, covering the period before the first practical class after the midterm exam, there is no experimental variation in the course. All students have access to the same version of the MicroApp and complete the same in-class surveys (which do not yet elaborate on weekly study goals and plans).⁶ Friday 5 October 2018 marks the transition between the first stage and the second stage. On this day with practical classes, the in-class surveys feature experimental

⁶The pre-experimental version of the MicroApp is shown in Appendix section 4.A.1. The standard format of the in-class survey used in the first stage is shown in Appendix section 4.A.2.

variation for the first time. The second stage begins on Saturday 6 October 2018 (start of lecture week 5), as experimental variation is now introduced in the MicroApp as well. That is, in the second stage the experiment is in full swing, with treatment-specific versions of the surveys and the platform.

The MicroApp was introduced to the students during the introductory lecture of the course, along with a few general comments that were discussed in Section 4.4.2. Personal login credentials were shared with students the next day via email. The actual recording of the time spent on the MicroApp (my measure of student effort) started a few days later in order to let it coincide with the start date of the regular lecture weeks in the course. I decided to do so because the introduction week (the week prior to the regular lecture weeks) is a non-standard week that is mainly reserved for information meetings, workshops and social events. These activities leave students with little time to engage in actual study activities.

Table 4.3: Timeline of the Experiment

Date	Event(s)
3 September 2018 - 4 October 2018	First stage (no experimental variation)
3 September 2018 - 7 September 2018	Introduction week at the university (start of the academic year)
4 September 2018	MicroApp introduced during an introductory lecture
5 September 2018	Personal MicroApp login credentials sent to students via email
10 September 2018	Start of regular lecture weeks in the course
	Start of recording the time spent on the MicroApp
14 September 2018	First in-class survey on time-use
	Extra questions in first survey on overall goals in the course
21 September 2018	Second in-class survey on time-use
28 September 2018	Third in-class survey on time-use
	Third survey includes a task with intertemporal choices about real effort
2 October 2018	<i>Midterm exam</i>
5 October 2018	Transition day. Fourth in-class survey on time-use
	The fourth and remaining surveys come in three versions
	Each experimental condition completes its own version
6 October 2018 - 1 November 2018	Second stage (experimental variation)
6 October 2018	From now on, three different MicroApp versions online
	Each experimental condition has access to its own version
12 October 2018	Fifth in-class survey on time-use
19 October 2018	Sixth in-class survey on time-use
26 October 2018	Seventh in-class survey on time-use
1 November 2018	<i>Final exam</i>

The surveys on time-use were completed in-class during the practicals on Friday. To ensure unique identification and a high response rate in the first stage, I attached the surveys to the (graded) weekly practical quizzes. Students first completed the quiz and afterwards were allocated extra time

to complete the survey. Students work on the quizzes and surveys individually (no talking allowed) and write down their name and/or student number once both are completed.⁷

In addition to the two core questions on time investments in learning inputs, the first-stage surveys were used to obtain some further information on the initial motivation of the students and the presence of self-control problems. In the first survey, I asked the students to write down their goals in terms of the overall course grade, grade on the midterm exam, and grade on the final exam. I also asked each student to indicate the chances of him or her participating in the final exam.⁸ The precise wording of the questions can be found in Appendix section 4.A.4. In the spirit of Augenblick, Niederle and Sprenger (2015), the third survey was in part used to administer a task that aims to measure whether students exhibit present biased preferences in real-effort decisions. The task closely follows the version used in Carvalho, Meier and Wang (2016). In the task, students have to indicate whether they wish to complete a long online research survey at a later date or a shorter version of the same survey at an earlier date. The question is asked repeatedly with variations in two dimensions: the length and the deadline of the shorter survey. The precise wording of the task can be found in Appendix section 4.A.5. The responses to these additional questions and tasks are discussed in Section 4.5.1.

Students were notified beforehand during the first lecture after the midterm exam that “some new features” are being tested during the second part of the course. They were made aware that this testing implies that different versions of the MicroApp and the in-class surveys are going to be used.⁹ Students were ensured that everyone would continue to have access

⁷The filled-in surveys were collected by teaching assistants and afterwards handed over to me for further processing.

⁸Asking a student to indicate the chances of him or her participating in the final exam serves as an indication for how committed the student is to complete the course. In the Netherlands, students sometimes take a course as a non-mandatory elective and in such cases the commitment to finish the course may not always be high.

⁹I slightly changed the procedure during the practicals in order to accommodate the experimental variation in the in-class surveys. I detached the surveys from the quizzes and instead placed them in three separate plastic folders. Each folder contained one version of the in-class survey as well as a list with students to whom the survey should be handed out. The lists follow from the randomization procedure in Section 4.4.4.

to the same course material on the MicroApp. It was also made explicitly clear to each student that whichever version is made available is completely determined by chance. Teaching assistants were trained on how to respond in case students had questions about the new procedure.¹⁰ This general information was provided to the students in order to avoid that the new procedure and the new versions come as a complete surprise.

The post-midterm survey responses are processed the same day in the MicroApp. For each student, I enter the MicroApp study goal for the upcoming week (all experimental conditions) and the study plan for the upcoming week (conditions II and III). The MicroApp automatically updates all accounts and versions with the new information. From the next day on, the information is displayed in the platform and students can at any time track their progress relative to their own goal and plan.¹¹ This procedure is repeated every lecture week until the final exam is administered. In case a student did not submit survey responses during a particular week, the information segments in the platform will remain empty.

4.4.6 Data collection

I operationalize study effort by focusing on the time spent on the MicroApp. The interventions of the experiment are geared toward helping students implementing their initial study plan with respect to this effort measure. Of course, I recognize that overall student effort is not limited to time spent on the MicroApp and may also involve time investments in other learning inputs, such as attending classes and reading the textbook. The issue with these other components of student effort, however, is that they are either difficult to measure (e.g., reading the textbook) or a mandatory aspect of

All students (panel A in Table 4.2) are included in the lists. The folders were shared with the student assistants prior to the start of each practical class. The class started with the quiz, followed by a short break. Afterwards, teaching assistants distributed the surveys by going through the lists and calling the student names. Students completed the surveys individually (no talking) and wrote down their name and/or student number on the survey. The latter made it possible to uniquely match survey responses with MicroApp accounts and administrative databases.

¹⁰In case of very specific questions or an unwillingness to further participate in the surveys, a student could at any time contact the secretary of the course. No student made use of this option.

¹¹In Section 4.4.3, I discussed how the information is displayed in the MicroApp.

the course (e.g., “participation grade” for class attendance). This makes these other measures of student effort less suitable for this study.

There are three reasons for focusing on the time spent on the MicroApp rather than, say, the number of solved practice tasks on the platform. First, time spent is a measure less confounded by ability. Second, effort on the MicroApp may also, among other things, entail reading feedback and/or recaps. This would not be well-captured by the alternative measure. Third, by focusing the study goal on the time spent on the platform, rather than the number of solved tasks, I avoid creating perverse incentives for skipping feedback and quickly solving many tasks by submitting random answers.

To verify whether the time spent on the MicroApp is used productively, fifteen minutes of inactivity (no mouse movements) will activate a pop-up requesting the student to respond by clicking on a button. In case of no response to this pop-up, the student will be automatically logged out after one minute and the counting of time will stop for that particular session. This procedure ensures that the measure of student effort captures active engagement with the course content on the platform. It also rules out that a student can meet his or her study goal by passively staying logged in.

To assess the relative magnitude of other learning inputs in students’ study efforts, the in-class surveys elicit weekly time investments in the following seven categories: (1) attending classes; (2) reading offline course material (e.g., textbook, lecture slides, hand-outs); (3) offline practice (e.g., tutorial exercises and old exams); (4) use of the MicroApp; (5) reading/watching online course material (not on the MicroApp); (6) online practice (not on the MicroApp); (7) any other study activity that is not captured by the previous categories (e.g., attending office hours). Students can tick per category one of the available answer boxes, varying from “not at all” to “five hours or more than five hours”.

To complete the dataset used in this study, I match the MicroApp data and the in-class survey responses with administrative databases that contain further student background information. Specifically, I obtain student-level data on gender, age, average grade in prior education, and the type of prior education. Through the administrative databases, I also collect information on class attendance and interim course performances.

4.5 Empirical evaluation

This section presents the empirical evaluation and discusses the findings. I start by performing balance tests on the first-stage data (baseline). The second stage (experimental period) is discussed afterwards in four different parts. In the first part, I examine how the interventions affect student effort (hypothesis I). I will distinguish between effort on the MicroApp and time investments in other learning inputs. The second part discusses goal completion rates (hypothesis II). Third, I cover the treatment effects on learning outcomes (hypothesis III). In the final part, I perform a more exploratory analysis that looks at the effects of learning inputs.

Throughout the discussion of the second stage, I will present the overall treatment effects as well as the treatment effects for the following three subgroups: males/females, high-performers/low-performers (based on pre-experimental data), and procrastinators/nonprocrastinators. These groups were prespecified in the pre-analysis plan of this study. I operationalize high/low pre-experimental performance as the within-treatment relative performance of a student on the midterm exam. Low-performance (high-performance) students are in the within-treatment bottom (top) 50% in terms of midterm performance.¹² Procrastinators are identified based on course enrollment dates (as will be discussed later).

4.5.1 First stage: no experimental variation

During the first stage of the study, no experimental variation is introduced. All students have access to the same version of the MicroApp and complete the same in-class surveys. Table 4.4 provides the first-stage descriptive statistics. Note that, as in the remainder of this chapter (unless stated otherwise), the table reports on the students that satisfy the inclusion rules of this study (covered in Section 4.4.4). The responses from the fourth in-

¹²In the pre-analysis plan (PAP), I operationalized initial ability as the average pre-university grade of a student. After completing the PAP, however, I was informed that this information is only available for the Dutch students in the sample. Using this operationalization would thus exclude all international students from the analysis (about 25% of all enrolled students). Furthermore, heterogeneity in the type of pre-university education hinders fair comparisons of pre-university grades. For these reasons, I will use the alternative operationalization discussed in the main text.

class survey (administered on 5 October 2018 (transition day)) are also included in this table. This is because the survey questions relate to the fourth lecture week (a first-stage week).

Students are on average about 20 years old at the start of the course. Most are in the first year of their study (70%). The share of female students in the sample is 30%. The majority of the students, about 75%, completed their pre-university education in the Netherlands. Restricting attention to the Dutch regular first-year students, as these share the same pre-university educational background, the average pre-university grade is 6.82 out of 10. Demographics appear balanced across the three experimental conditions.

Students are in general committed to completing the course and rate the chances of participating in the final exam at roughly 97%. The students aim for an overall course grade of 7.69 (out of 10). Performance goals are similar for the midterm exam (7.94) and the final exam (7.61). The initial course goals of the students are similar across the three conditions.

Focusing on pre-experimental behavior on the MicroApp, I again observe that students in the different conditions behave in the same way. Students on average spent six hours studying on the MicroApp during the first stage. Each student solved about 70 practice tasks, of which 62% was answered correctly. A little less than three minutes was spent on each task, which suggests that students at least took the time to read the question and did not simply click through in an inattentive manner. Students loaded recap modules (clicked on the start button) about eighteen times on average, which is a rough indication of having read the respective recaps.

Performances in other aspects of the course display the same pattern. Students scored an average of 21 points (out of 30) on the midterm exam, which translates in a grade of 7 (out of 10). Similar results are found for the weekly practical quizzes, on which the students scored a 7.5 (out of 10). Class attendance was high during the first stage, which indicates high involvement in the course. Although this may in part be due to attendance being incentivized through a participation grade.

Considering the pre-experimental responses to the in-class surveys, we get an indication of how students spent their time in the first part of the course. Per learning input students could tick one of the five available an-

Table 4.4: First-Stage Descriptive Statistics of Experimental Conditions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Full sample		T1: Control		T2: Planner		T3: Rabbit		Joint test: treatment effect = 0	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.		
Pre-experimental behavior on the MicroApp										
Avg. time on the app (in hours)	6.14	(6.43)	6.04	(6.57)	6.07	(6.42)	6.30	(6.36)	0.05	[0.95]
Number of solved tasks	69.49	(61.20)	70.69	(63.59)	70.48	(61.38)	67.37	(59.12)	0.10	[0.91]
Time needed to solve a task (in minutes)	2.76	(1.73)	2.62	(1.59)	2.87	(1.91)	2.79	(1.68)	0.58	[0.56]
Share of tasks solved correctly (%)	61.74		61.36		62.44		61.44		0.14	[0.87]
Recap activities	18.40	(16.74)	18.42	(16.60)	17.35	(15.94)	19.38	(17.71)	0.34	[0.71]
Pre-experimental behavior in the course										
Grade on practical quizzes (out of 10)	7.50	(1.04)	7.47	(1.03)	7.50	(0.98)	7.53	(1.11)	0.07	[0.93]
Score on the midterm exam (out of 30)	21.02	(5.21)	21.28	(5.32)	20.92	(5.54)	20.86	(4.80)	0.20	[0.82]
Number of tutorials and practicals attended	7.43	(0.84)	7.45	(0.81)	7.37	(0.90)	7.46	(0.81)	0.31	[0.73]
Demographics										
Age at the start of the course (in years)	20.49	(2.23)	20.30	(2.17)	20.65	(2.12)	20.53	(2.40)	0.66	[0.52]
Avg. pre-university grade (out of 10)	6.82	(0.53)	6.80	(0.55)	6.86	(0.55)	6.80	(0.50)	0.23	[0.80]
Share of female students (%)	30.26		33.66		30.61		26.67		0.60	[0.55]
Study phase (# students)										
Regular first-year students	216		73		68		75		0.05	[0.95]
Minor finance students	43		15		14		14		0.06	[0.94]
Pre-MSc students	50		16		17		17		0.04	[0.96]
Initial course goals of students										
Overall course grade (out of 10)	7.69	(0.88)	7.77	(0.94)	7.68	(0.91)	7.62	(0.80)	0.56	[0.57]
Midterm exam (out of 10)	7.94	(0.95)	8.02	(0.90)	7.86	(1.04)	7.95	(0.89)	0.53	[0.59]
Final exam (out of 10)	7.61	(1.03)	7.76	(1.09)	7.59	(0.97)	7.51	(1.01)	1.13	[0.32]
Chances of taking the final exam (out of 100%)	96.64		97.93		96.90		95.32		1.50	[0.22]
Pre-experimental median responses to in-class surveys										
Time spent on course learning inputs (per week)										
Attending classes	5		5		5		4		0.81	[0.45]
Reading offline material	3		3		3		3		0.33	[0.72]
Practicing with offline material	3		3		3		2		5.01	[0.01]
Use of the MicroApp	2		2		2		2		0.15	[0.86]
Reading/watching online material (not MicroApp)	1		1		1		1		1.58	[0.21]
Practicing with online material (not MicroApp)	1		1		1		1		0.57	[0.57]
Other (not captured by the above)	1		1		1		1		5.57	[0.00]
Avg. time spent on other courses (per week)										
Other courses	3		3		3		3		2.31	[0.10]
Number of students	309		104		99		106			

Notes: Unit of observation is the student. Columns (9) and (10) show F -statistics and p -values from a balance test of whether the treatment coefficients are jointly equal to zero. Students' demographics and study phase are retrieved from administration records of the university. The average pre-university grade is based on data from Dutch regular first-year students, as these students have the same pre-university educational background. Minor finance students and Pre-MSc students are senior students who follow a tailored deficiency program (of which the microeconomics course is a mandatory component) before they are allowed to enroll in a MSc program at the university (see footnote 3 in this chapter for further details). Initial course goals of the students are collected during the first in-class survey in the course. With regard to the survey question related to the weekly time investments in course learning inputs, students could tick one of the following available answer boxes: (1) not at all, (2) less than 1 hour, (3) 1 – < 3 hour(s), (4) 3 – < 5 hours, (5) 5 hours or more than 5 hours. The question on the average weekly time investment in other courses had the following answer boxes: (1) not at all, (2) less than 5 hours, (3) 5 – < 10 hour(s), (4) 10 – < 15 hours, (5) 15 hours or more than 15 hours.

swer boxes with weekly time investments, varying from *not at all* to *5 hours or more than 5 hours*.¹³ Table 4.4 shows that across all conditions the most popular learning inputs are attending classes, reading offline, practicing offline, and use of the MicroApp.

Figure 4.4 aggregates per lecture week the first-stage time investments in learning inputs. This figure reveals that time investments change during the first-stage lecture weeks. Early on in the course, students spent most of their time “consuming” course material (e.g., attending classes and reading), whereas later on the focus shifts toward practicing. Practicing is mainly done with offline material (e.g., old exams) or online on the MicroApp. Few students indicate that they make use of other online resources. The surveys also show that the average time investment in other courses is similar across the three groups (5-10 hours per course).¹⁴

Overall, the previous discussion indicates that students in all three experimental conditions are similar in terms of observable characteristics in the pre-experimental period. This suggests that the randomization is successful in creating balance before experimental variation is introduced, and that this result is robust to the applied exclusion rules.

Do we observe present-biased behavior among students prior to the start of the experiment? In one pre-experimental in-class survey, students completed an incentivized multiple price list with intertemporal choices about filling in a shorter online survey earlier or a longer online survey later.¹⁵ Note that these choices relate to real effort rather than monetary rewards. This is for two reasons. First, it more closely resembles the decision problem of a student in which I am interested, namely the decision how much study effort to exert at any given time. Second, intertemporal choices about mon-

¹³The full choice set was as follows: (1) not at all, (2) less than 1 hour, (3) 1– < 3 hour(s), (4) 3– < 5 hours, (5) 5 hours or more than 5 hours.

¹⁴For the survey question related to the average time investment in other courses, the available answer boxes were: (1) not at all, (2) less than 5 hours, (3) 5– < 10 hour(s), (4) 10– < 15 hours, (5) 15 hours or more than 15 hours.

¹⁵The multiple price list can be found in Appendix section 4.A.5. The online survey was an actual survey from a research team at the university (unaffiliated with this study). In line with the information stated in the multiple price list, there was a short version of the online survey and a long version. The short version, in turn, varied in length such that it would take approximately 15 minutes, 18 minutes, 21 minutes, 24 minutes or 27 minutes to complete.

Figure 4.4: First-Stage Weekly Time Investments in Learning Inputs

(a) Lecture week 1

(b) Lecture week 2

(c) Lecture week 3

(d) Lecture week 4

Notes: Students' time investments in learning inputs are elicited at the end of each lecture week via in-class surveys. In the first stage of the experiment (overlapping with the first four lecture weeks) there is no experimental variation, which means that all students have access to the same version of the MicroApp and the in-class surveys. Halfway through lecture week 4 students participated in a mid-term exam (worth 30% of the final grade). The learning input “classes” is partly incentivized (attendance is required to be eligible for a participation grade (worth 10% of the final grade)). The MicroApp is a newly introduced online learning platform in the course with abundant practice- and recap material. The categories “read/watch online” and “practice online” refer to non-MicroApp online study activities.

etary rewards may not be well-suited to capture dynamic time-inconsistent preferences due to several confounds (such as payment reliability). After controlling for these confounds, recent studies find no or limited evidence of time inconsistency in monetary choices (Giné, Goldberg, Silverman, and Yang 2018, Andreoni and Sprenger 2012, Andersen, Harrison, Lau, and Rutström 2014). Augenblick et al. (2015) also report a lack of evidence of time inconsistency in monetary decisions, but they do find considerable present bias in decisions by students that are based on real effort.

In the multiple price list method, I closely follow Carvalho et al. (2016) and let students choose between soon completing a short survey or later completing a long survey, where I vary the length of the shorter survey. The choices are presented to students in two time frames: a short time frame (5 days (sooner) x 35 days (later)) and a long time frame (90 days (sooner) x 120 days (later)). There are ten choices in total, five per time frame. Similar to the approach in Meier and Sprenger (2015), I elicit individual discount rates (IDRs) by looking at the point where a student switches from the short survey to the long survey in each time frame. These switching points yield interval information about students' intertemporal preferences.¹⁶

In Table 4.5, I present the results of interval regressions where the dependent variable is an interval measure of the individual discount rate (IDR). As explanatory variables I include indicators for the short time frame and treatment status. Including treatment status serves as an additional balance check, since IDRs were elicited before experimental variation was introduced. Present bias in real effort is observed when the individual discount rate is higher in the short time frame than in the long time frame. In column (1) we observe that the point estimate of the IDR is indeed higher in the short time frame, although it is not different from zero at conventional significance levels. As expected, treatment status has no impact on IDRs. Results are similar when I only look at students who made

¹⁶As in Carvalho et al. (2016), let x be the duration in minutes of the longest sooner survey a student chooses over the later survey. The (lb, ub) (lower bound, upper bound) of the discount rate intervals are then: $(15/30, 18/30)$ for $x = 15$; $(18/30, 21/30)$ for $x = 18$; $(21/30, 24/30)$ for $x = 21$; $(24/30, 27/30)$ for $x = 24$, and $(27/30, 30/30)$ for $x = 27$. For those students who always pick the later survey, I will use the interval $(0, 15/30)$. For each time frame, I respectively calculate the lower bound and upper bound of the individual discount rates (IDRs) as follows: $IDR = \frac{1}{ub} - 1$ and $IDR = \frac{1}{lb} - 1$.

Table 4.5: Pre-Experimental Intertemporal Choices About Real Effort

	(1)	(2)	(3)	(4)
Short time frame=1	0.041 (0.046)	0.056 (0.042)	0.047 (0.035)	0.048 (0.032)
Treatment=2	-0.053 (0.058)	-0.040 (0.052)	-0.030 (0.048)	-0.028 (0.043)
Treatment=3	0.087 (0.065)	0.074 (0.061)	0.078 (0.053)	0.050 (0.049)
Short time frame=1 $\times$ Treatment=2	0.093 (0.067)	0.082 (0.064)	0.065 (0.051)	0.073 (0.048)
Short time frame=1 $\times$ Treatment=3	-0.074 (0.067)	-0.041 (0.060)	-0.061 (0.055)	0.005 (0.049)
Constant	0.334*** (0.044)	0.259*** (0.041)	0.319*** (0.036)	0.254*** (0.034)
Number of students	220	188	333	286
Number of observations	440	376	666	572
Student inclusion/exclusion rules applied?	Yes	Yes	No	No
Conditional on consistent choices?	No	Yes	No	Yes

Notes: This table reports the estimates of interval regressions where the dependent variable is an interval measure of the individual discount rate (IDR). Standard errors are clustered at the student level. Students were asked to make intertemporal choices about completing a shorter survey earlier or a longer survey later. Two time frames were presented to the students: a short time frame (5 days (sooner) $\times$ 35 days (later)) and a long time frame (90 days (sooner) $\times$ 120 days (later)). IDRs are estimated for each time frame. Only students who completed all choices in the price list are considered. Choices are consistent when a student has at most one switching point in each time frame. ***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

consistent choices (about 85% of the students). I consider choices to be consistent when there is at most one switching point per time frame (which satisfies monotonicity). There is no change in outcomes when the full sample of students is considered (no exclusion rules applied). In all cases, the point estimate of the IDR is higher in the short time frame (about 5 percentage points higher), suggesting present bias, but the estimate is never significantly different from zero.

The two time frames make it possible to identify per student whether they exhibit present bias in their intertemporal effort choices. That is, when their IDR is higher in the short time frame than in the long time frame. Let K_{ST} (K_{LT}) be the switching point in the short time frame (long time frame). I classify a student as present-biased in effort when $K_{LT} > K_{ST}$. Using this definition, about 10% of the students display present bias in their choices. This reduces to 8% when only consistent choices are considered.¹⁷

In the pre-analysis plan, I indicated that I am interested in performing a subgroup analysis by contrasting outcomes between present-biased students and non-present-biased students. The multiple price list method, however, yields a sample of present-biased students that is too small for conducting

¹⁷Percentages are not much different when the full group of students is evaluated.

a meaningful analysis. The small number identified may in part be due to difficulties in task comprehension and the fact that many students had no switching point at all.¹⁸

An alternative way of identifying students prone to procrastination is proposed by Himmeler, Jäckle and Weinschenk (2019). They use the median application date to an educational program as a cutoff for distinguishing procrastinators (applied after the median date) from nonprocrastinators (applied before), reasoning that nonprocrastinators are more likely to take care of things right away (such as the application procedure) and not wait until the last minute. The application date is treated as a proxy for procrastination tendencies.¹⁹ Just like the multiple price list method, a benefit of this approach is that it relies on actual behavior and not on self-reported measures (which may be flawed when answered in a socially-desirable way).

I will adopt a similar approach by looking at the course enrollment dates of students. Figure 4.A.9 shows the number of students per enrollment date.²⁰ The most popular day for enrolling is the day of the registration deadline, which is suggestive of procrastination. The median course enrollment date in the full sample (21 August 2018) is a little less than one week before the registration deadline (26 August 2018). About 30% of the students registered after the deadline. These students were still granted access to the course and were given permission to take the exams.²¹

The within-treatment median course enrollment date is used as the cutoff for identifying procrastinators and nonprocrastinators in each treatment. Table 4.6 summarizes the results per treatment. The random allocation of

¹⁸After filling in the in-class survey, some students indicated to teaching assistants that they did not fully understand the survey part related to the multiple price list. About 32% of the students had no switching point at all. That is, in all ten choices these students always chose the same version of the online survey.

¹⁹Other recent papers also make use of enrollment/application behavior to gauge procrastination tendencies. Reuben, Sapienza and Zingales (2015) consider the application date to an MBA program as a procrastination proxy. De Paola and Scoppa (2015) look at the time elapsed between receiving a university acceptance letter and the date of enrolling. Brown and Previtero (2014) study financial behaviors and define a procrastinator as an individual who waits until the last day of the enrollment period before making a decision which health care plan to select.

²⁰The figure shows all enrolled students in the course. No exclusion rules are applied.

²¹A further check reveals that indeed 30% of the students who took the midterm exam belongs to the group of students that registered late.

students to the experimental conditions successfully balances the median enrollment date in each group. In total, 268 students (289 students) are identified as procrastinators (nonprocrastinators).

Table 4.6: Identifying Procrastinators in the Course

	Median enrollment date ($\bar{D}$)	# Students (% share)	
		Procrastinators ($> \bar{D}$)	Nonprocrastinators ($\leq \bar{D}$)
All students	21 August 2018	268 (48.11%)	289 (51.89%)
Condition I: control condition	21 August 2018	84 (46.65%)	100 (54.35%)
Condition II: “planner” condition	20 August 2018	92 (49.20%)	95 (50.80%)
Condition III: “rabbit” condition	21 August 2018	92 (49.46%)	94 (50.54%)

Notes: The within-treatment median course enrollment date is used as a cutoff for identifying procrastinators and nonprocrastinators in each treatment. All enrolled students are considered (no inclusion/exclusion rules are applied). For 16 students the enrollment date could not be identified.

Some basic analyses (not shown) reveal that students identified as procrastinators perform slightly worse in the course prior to the experimental variation. They score, on average, about 0.6 points lower on the midterm exam (a total of 30 points could be obtained). Himmeler, Jäckle and Weinschenk (2019) also observe that procrastinators perform worse in exams. They find that procrastinating students tend to pass fewer exams.

4.5.2 Second stage: experimental variation

Time investments in learning inputs The first hypothesis I wish to examine, as detailed in Section 4.3, is that students who observe their own pacesetter will exert higher study effort than students who do not observe their own pacesetter. I find support for this hypothesis when students in the rabbit condition invest significantly more time in the learning input targeted by the treatments (studying on the MicroApp). I will start by discussing study effort on the MicroApp and then continue with students’ time investments in the other learning inputs.

Figure 4.5 provides an overview of daily student effort (aggregated) on the MicroApp during the course. Panel A (panel B) of the figure displays the sum total of daily study time (solved practice tasks) per experimental condition. Both measures of student effort display the same overall pattern. In the first stage (the period before the fourth practical class (P_4)), students in all three experimental conditions bunch their effort right before the due date of a graded course component, such as practical quizzes and the midterm exam. That is, overall student activity is much higher in

Figure 4.5: Sum Total of Daily Study Time on the MicroApp

(a) Study time

(b) Solved practice tasks (alternative measure)

Notes: The figure shows per experimental condition the sum total of study time (panel A) and the number of solved practice tasks (panel B) on the MicroApp. The regular lecture weeks of the course started on Monday 10 September 2018. Surveys on time-use were administered during each practical class (P). Experimental variation in the surveys was introduced during the fourth practical class ($P4$) on 5 October (the first practical after the midterm exam) and continued until the final exam. Three different MicroApp versions (one per experimental condition) aired on 6 October and stayed online until the final exam. The period before (after) 5 October marks the first (second) stage of the experiment.

the days right before such events. This *may* be indicative of self-control problems on behalf of the students. Students may therefore see potential in using the planning tools provided in the interventions to get started earlier on preparing for the graded tasks. Consistent with the discussion in Section 4.5.1, I observe no significant behavioral differences among the three experimental groups in the first-stage period.

In the second stage (the period after the fourth practical class ($P4$)), the interventions are in place and used by the students. That is, each week students are asked to formulate a study goal for the upcoming week (all conditions) and to indicate via a detailed study plan how this goal will be reached (planner- and rabbit condition). This study plan, in turn, is used to construct the personalized pacesetter (rabbit condition). At first sight, the interventions do little in terms of inducing students to start earlier with studying. All conditions display the same pattern of student effort, which is similar to the pattern observed in the first stage. Most students continue to bunch their effort right before the due dates of graded tasks. I will later elaborate on the question to what extent this observed study behavior is consistent with the initial study plans of students.

Figure 4.6 examines whether the interventions increase student effort during the week. The figure plots per experimental condition the average weekly study time on the MicroApp, conditional on having logged in at least once during the week. Study time during the transition day between the first- and the second stage is excluded from the analysis. The figure shows that average student effort starts off low in the beginning of the course but gradually increases once exams are approaching. In the first stage (weeks 1-4), average student effort appears to be the same across the three conditions. In the second stage (weeks 5-8), students in the control condition (T1) and the rabbit condition (T3) continue to study at more or less the same pace. Students in the planner condition (T2), however, seem to lag behind in terms of study effort, especially in the exam week.

To formally identify the treatment effects δ on weekly study time, I adopt a differences-in-differences regression specification that models the MicroApp study time Y_{it} (in minutes) of student i in week t conditional on treatment status T_i and an indicator for the weeks with the interventions

Figure 4.6: Average Weekly Study Time on the MicroApp

Notes: The figure shows per experimental condition the average weekly study time on the MicroApp, conditional on having logged in at least once during the week. Experimental variation is introduced in the in-class surveys during the fourth practical class (at the end of lecture week 4). Study time during this transition day is excluded from the analysis. Three different MicroApp versions (one per experimental condition) aired the next day (start of week 5) and stayed online until the final exam.

fully implemented (from week 5 on). I refer to this indicator as $post_t$. Student fixed effects μ_i are also included to control for time-invariant student-specific characteristics that may affect study time, such as initial ability. The regression model is estimated with robust standard errors, clustered at the student level so as to account for within-student correlation patterns in the error term ϵ_{it} (Bertrand, Duflo, and Mullainathan 2004). The model specification is as follows:

$$Y_{it} = \beta \cdot post_t + \sum_{j=1}^3 I\{T_i = j\} \cdot (\delta_j \cdot post_t) + \mu_i + \epsilon_{it}. \quad (4.1)$$

Table 4.7 reports the results. The estimates in the table can be best interpreted as treatment-on-the-treated (ToT) effects, since students are only

Table 4.7: Treatment Effects on MicroApp Study Time

	(1)	(2)	(3)	(4)	(5)
Post=1	106.007*** (23.141)	119.549*** (21.338)	102.288*** (22.082)	109.408*** (19.850)	0.678*** (0.151)
Post=1 $\times$ Treatment=2	-60.906** (30.906)	-66.413** (28.164)	-51.392* (29.477)	-44.109* (26.635)	-0.435* (0.239)
Post=1 $\times$ Treatment=3	-46.587 (32.563)	-32.771 (30.307)	-36.611 (31.657)	-25.937 (28.481)	-0.364 (0.238)
Treatment=2		-1.165 (22.629)		2.950 (20.956)	
Treatment=3		22.294 (22.894)		21.872 (21.077)	
Constant	195.461*** (6.288)	180.656*** (15.189)	185.166*** (6.362)	169.808*** (14.098)	4.265*** (0.049)
R ²	.0303	.0339	.0276	.03	.025
Number of students	305	305	413	413	305
Number of observations	1,174	1,174	1,376	1,376	1,174
Student fixed effects	Yes	No	Yes	No	Yes
Student inclusion/exclusion rules applied?	Yes	Yes	No	No	Yes
Time unit	Week	Week	Week	Week	Week
Dependent variable	Level	Level	Level	Level	Log

Notes: Identification of the treatment-on-the-treated effects on MicroApp study time (in minutes). Robust standard errors (clustered at the student level) are in parentheses. The variable *post* captures the experimental period. Study time during the transition day (5 October 2018) from the baseline period (lecture weeks 1-4) to the experimental period (lecture weeks 5-8) is excluded. There are two exclusion rules. First, a student is excluded when he or she participated in less than two practicals after the midterm exam. Second, a student is also excluded when he or she never made use of the MicroApp in the period until the first practical after the midterm exam. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

included when they logged in at least once during the week. My interest is mainly in the behavior of these students because they have been exposed to the treatment variation on the MicroApp and have at least attempted to exert non-zero study effort. Intention-to-treat (ITT) effects are presented in Appendix section 4.A.7 and are qualitatively similar.²²

Consistent with the interpretation of Figure 4.6, I find some (weak) support for the observation that students in the planner condition lag behind in terms of weekly study time on the MicroApp. Planner students study about one hour less per week compared to control students. This negative effect extends to the full sample of students (when no inclusion/exclusion rules are applied). Table 4.7 suggests that the pacesetter feature does not increase study time on top of what is already observed among control students in the experimental period. If anything, the pacesetter has a negative impact on study time as well, although these effects are not significantly different from zero at conventional significance levels.

²²The ITT estimates are based on all students, including those with zero study effort on the MicroApp during the week. While the ITT point estimates of study time are lower in the different specifications, the results are qualitatively similar to the ToT estimates. As a robustness check, I also estimated the ITT effects with Tobit regressions and found similar results. These results are available upon request.

With the available data, I therefore find no support for the first hypothesis in Section 4.3 in terms of overall treatment effects. Students who can observe their own pacesetter do not study more than students who cannot. Later I will examine study commitments more closely and discuss whether the treatments succeeded in helping students to reach their goals.

The absence of an overall treatment effect need not imply that there are no temporal effects of the interventions. Given the multi-week experimental period and the multiple rounds of setting study plans and pacesetters, this can be a real possibility. For example, the interventions may have had an impact only in the beginning of the experimental period. The richness of the data makes it possible to explore such temporal effects. Figure 4.7 shows the weekly treatment effects associated with the two interventions. The figure plots coefficients $\gamma_{j\tau}$ and $\delta_{j\tau}$ from the following regression specification:

$$Y_{i\tau} = \sum_{j=1}^3 \sum_{\tau=m+1}^{-1} \gamma_{j\tau} D_{ij\tau} + \sum_{j=1}^3 \sum_{\tau=0}^q \delta_{j\tau} D_{ij\tau} + v_{\tau} + \mu_i + \epsilon_{i\tau}. \quad (4.2)$$

Where τ captures the proximity in weeks from the fifth lecture week (the implementation week), with $\tau \in [m, q]$. Note that m is the first lecture week ($\tau = -4$) and q is the exam week ($\tau = 3$). The implementation week (start of the experimental period) is $\tau = 0$. The leads (lags) are therefore captured by $\tau < 0$ ($\tau \geq 0$). Interactions between the treatment indicator T_i and the time dummies v_{τ} are captured by $D_{ij\tau}$. As in Equation 4.1, the model is estimated with student fixed effects μ_i and robust standard errors (clustered at the student level). The first week is treated as the baseline.

The figure reveals no immediate effects of the treatments, nor do I in general observe delayed responses. At the end of the experimental period, close to the final exam, treated students appear to study slightly less than control students (at least on the MicroApp). In line with the results from Table 4.7, however, the overall impression is that the treatments do little in terms of inducing students to exert extra study effort. The figure also shows that the coefficients of the leads, $\gamma_{j\tau}$, are not significantly different from zero. This suggests pre-experimental balance and no anticipation effects.

Figure 4.7: Temporal Treatment Effects on MicroApp Study Time

(a) Condition II: Planner

(b) Condition III: Rabbit

Notes: The figure plots coefficients $\gamma_{j\tau}$ and $\delta_{j\tau}$ from Equation 4.2. Vertical spikes indicate 95% confidence intervals. The first lecture week is taken as the reference week. The fifth lecture week is the implementation week of the experiment (point 0 on the x -axis). The dependent variable is the weekly study time (in minutes) of a student on the MicroApp. Regressions are estimated with robust standard errors (clustered at the student level) and include student- and time fixed effects. Only students who have logged in at least once during the week are included in the sample. Study time during the transition day from the baseline period to the experimental period is excluded.

Further analyses indicate that there are no interaction effects of the treatments with gender, pre-experimental course performance, and prior procrastination tendencies. In Table 4.A.2, I separate the effects for males and females. The signs suggest that the pacesetter feature induces male students to study less compared to their counterparts in the control group. The negative effect is less pronounced among female students, but the difference with males is not significant. This pattern is consistently observed across the different model specifications. In Table 4.A.3, I compare high-performance students with low-performance students (based on pre-experimental performance in the course) and observe that there is no treatment heterogeneity between these two groups. Students with procrastination tendencies also do not respond differently to the treatments compared to nonprocrastinating students, as is shown in Table 4.A.4.

The previous analyses highlight that the two treatments are ineffective in increasing study effort among students. Importantly, though, this only concerns study effort on the MicroApp, as this is the learning input that was targeted by the treatments (for reasons discussed in Section 4.4.6). Thus far, I have not considered the possibility that the treatments affected student effort indirectly through effort in other learning inputs.

Figure 4.A.10 in the appendix aggregates the in-class survey responses per lecture week in the experimental period.²³ There were no classes during the exam week and hence no survey responses are available for that week. The survey responses from the lecture weeks closely mirror the results from the baseline period. Most of the students' time is devoted to attending classes, followed by reading/practicing offline and use of the MicroApp. Other online resources appear to be rarely used by students.

In Table 4.8, I examine whether the treatments had an impact on time investments in the other learning inputs. The table shows the estimates based on interval regressions where the dependent variable is an interval

²³Note that the second-stage survey responses show that a large share of students indicated that they spent on average less than one hour per week on the MicroApp. At first sight, this may not seem in line with Figure 4.6. The difference is that in Figure 4.A.10 the responses are shown of all students, including the ones who never used the MicroApp during a specific week. Figure 4.6 only includes those students who used the MicroApp at least once during a specific week.

Table 4.8: Treatment Effects on Time Investments in Other Learning Inputs

Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Classes	Offline Reading	Offline Practice	Online Read/watch	Online Practice	Other	Other Courses
Post=1	-0.738*** (0.126)	-0.250** (0.110)	-0.503*** (0.094)	-0.123*** (0.046)	-0.121** (0.049)	-0.182*** (0.047)	0.246 (0.338)
Treatment=2	-0.140 (0.242)	-0.153 (0.191)	-0.161 (0.138)	-0.049 (0.083)	-0.077 (0.066)	-0.254*** (0.074)	-1.053* (0.557)
Treatment=3	-0.237 (0.237)	-0.200 (0.187)	-0.409*** (0.135)	0.064 (0.082)	-0.015 (0.065)	-0.156** (0.072)	-1.451*** (0.546)
Post=1 × Treatment=2	0.012 (0.177)	-0.042 (0.156)	-0.008 (0.133)	0.052 (0.065)	0.012 (0.069)	0.182*** (0.067)	0.383 (0.481)
Post=1 × Treatment=3	-0.109 (0.174)	-0.382** (0.153)	0.119 (0.131)	0.041 (0.065)	0.074 (0.069)	0.159** (0.067)	0.432 (0.467)
Constant	5.098*** (0.171)	2.678*** (0.133)	1.921*** (0.096)	0.726*** (0.058)	0.703*** (0.046)	0.812*** (0.051)	10.276*** (0.390)
Number of students	309	309	309	309	309	309	308
Number of observations	1,926	1,922	1,920	1,923	1,926	1,921	1,614
Student inclusion/exclusion rules applied?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time unit	Week	Week	Week	Week	Week	Week	Week

Notes: Identification of the treatment effects on time investments in learning inputs other than the MicroApp. Estimates are based on interval regressions where the dependent variable is a time interval. At the end of each lecture week, students were asked through in-class surveys to report their weekly time investments in various learning inputs. The following four time intervals are used in the regressions: $[0, 1)$, $[1, 3)$, $[3, 5)$, $[5, \infty)$, where the numbers are in hours. For the dependent variable *other courses* (average time investment in other courses), the following four time intervals are used: $[0, 5)$, $[5, 10)$, $[10, 15)$, $[15, \infty)$, where the numbers are again in hours. The variable *post* captures the experimental period. No in-class surveys were administered at the end of the exam week. The learning input “classes” is partly incentivized (attendance is required to be eligible for a participation grade (worth 10% of the final grade)). The learning inputs “online read/watch” and “online practice” refer to non-MicroApp online study activities. ***(**; *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

measure of the time investment in a learning input. The weekly survey responses enable me to determine for each student the appropriate time interval that best represents his or her time investment in a learning input during a specific week. Furthermore, given that the in-class surveys were administered during both the baseline period and the experimental period, I can compare differences between the three conditions before and after experimental variation is introduced.

The table corroborates the previous finding that the treatments did not significantly alter study effort among students, nor did it lead to substitution between learning inputs. Prior to the experiment, students in the rabbit condition appear to engage in slightly less practicing offline and also invest less time in other courses. During the experimental period, however, I find no evidence that treated students engage differently with non-MicroApp learning inputs. Hence, the treatments seem to have had no impact on studying.

Study goals and goal completion rates Next, I discuss the treatment effects on study goal completion rates. The ex-ante reasoning was that both the planner and the rabbit condition would make progress toward a student’s initial study goal more salient. The weekly study goal is transformed into a study plan with concrete daily goals that help the student to stay on track. These initial study commitments serve as reference points and induce the student not to postpone studying, thereby making the student less susceptible to self-control problems. I would thus expect that the interventions better enable students to reach their initial study goals.

Table 4.9: Percentage Shares of Non-Zero Weekly Study Goals

	T1: Control	T2: Planner				T3: Rabbit			
	%	%	Δ (T2-T1)	χ^2	p -value	%	Δ (T3-T1)	χ^2	p -value
All experimental weeks	80.27	82.30	2.03	0.49	0.485	83.51	3.24	1.31	0.252
Lecture week 5	82.98	87.91	4.93	0.90	0.342	88.89	5.91	1.40	0.237
Lecture week 6	80.41	81.72	1.31	0.05	0.818	78.43	-1.98	0.12	0.730
Lecture week 7	77.78	75.00	-2.78	0.21	0.648	79.38	1.60	0.07	0.784
Exam week	80.00	85.53	5.53	0.81	0.368	88.46	8.46	2.07	0.150

Notes: The table reports per experimental condition the percentage shares of students who have chosen a non-zero weekly study goal. The study goal is about how many hours students plan to study on the MicroApp in each week. Reported p -values are based on pairwise χ^2 -tests between the control and treatment conditions.

Table 4.9 shows that most of the students choose to set non-zero weekly study goals ($\approx 82\%$). The percentage shares are similar across the experimental conditions and do not vary much throughout the experimental period. I will restrict attention to these students with non-zero study goals, so as to avoid that goal completion rates are affected by students who trivially meet their goal by setting it equal to zero.²⁴

A closer look at the weekly study goals reveals that students in both treatment conditions set significantly higher goals, as shown in Figure 4.8. Considering the full experimental period, pairwise Wilcoxon-Mann-Whitney (WMW) tests show that the differences with the control group are significant for both the planner treatment ($p < 0.001$) and the rabbit treatment ($p < 0.001$). The pacesetter, for example, induces students to set goals that are about twice as high as those set by students in the control group.

The pattern among treated students of setting higher study goals is observed for both male and female students. Figure 4.A.11 reports the average study goal set by students in the experimental period, split by gender and treatment status. The figure shows clear within-gender differences between the different treatments. Both males and females in the planner and rabbit conditions set higher goals than their counterparts in the control condition.²⁵ Within treatments, I observe no significant differences between males and females.²⁶ Hence, irrespective of gender, the treatments induce students to set more ambitious study goals, with no evidence that the effect is stronger or weaker for a particular gender group.

Similar results are obtained when a distinction is made between students who performed well or poor in the pre-experimental period of the course. Figure 4.A.12 shows again that treated students set higher study goals,

²⁴I also exclude weekly study goals that are unrealistically high (defined as five SDs above the mean). This leads to the exclusion of seven observations (0.78% of the sample). These students indicated that they plan to study more than 40 hours on the MicroApp in a week (which would be similar in hours to a full-time lecture week). Not excluding these observations does not alter the main results: the p -values mentioned later in the main text are still $p < 0.001$ (based on pairwise WMW tests).

²⁵Pairwise WMW tests yield the following p -values for male students: $p < 0.001$ (T2: planner) and $p < 0.001$ (T3: rabbit). For female students, the results are: $p = 0.002$ (T2: planner) and $p < 0.001$ (T3: rabbit).

²⁶The p -values based on WMW tests are: $p = 0.545$ (T1: control), $p = 0.795$ (T2: planner), $p = 0.955$ (T3: rabbit).

Figure 4.8: Weekly Study Goals for MicroApp Usage

Notes: The figure shows per experimental condition how many hours students on average plan to study on the MicroApp in each week, conditional on having set a non-zero study goal. Weekly study goals are elicited each week through in-class surveys. The error bars denote ± 1 standard error.

regardless of whether they performed well or poor before the interventions were in place.²⁷ Interestingly, only in the rabbit condition do I observe clear differences in goal-setting between high-performance students and low-performance students.²⁸ The latter group on average sets higher study goals (6.8 hours versus 4.8 hours). This difference is not observed in the planner condition, suggesting that it is the pacesetter feature that drives this result and not the act of formulating a study plan.

In Figure 4.A.13, I show per experimental condition how procrastinating students differ from nonprocrastinating students in terms of their weekly study goals. As with gender and prior course performance, I observe once again that treated students set higher study goals, especially among those identified as nonprocrastinators.²⁹ Within the planner condition and the rabbit condition, the figure suggests that procrastinating students set lower study goals than nonprocrastinating students, although the effects are only marginally significant.³⁰

A benefit of the MicroApp and the elicited study plans of students is that I can trace in real time how differences in planned and actual study time develop. In Figure 4.9, the green dot (red triangle) depicts planned (actual) daily study time on the MicroApp, conditional on having logged in at least once during the week and having set a non-zero weekly study goal. The bars show the daily difference between actual and planned study time, with a negative value indicating that students on average fell short of their own study goal of that day. In general, students in both treatment conditions are too optimistic when planning their daily study effort; planned study time is in most cases higher than actual study time. The figure also shows that both planned and actual effort increase in the days before the practical quizzes and the final exam, which suggests that effort bunching is

²⁷Pairwise WMW tests yield the following p -values for high-performing students: $p < 0.001$ (T2: planner) and $p < 0.001$ (T3: rabbit). For low-performing students, the results are: $p = 0.009$ (T2: planner) and $p < 0.001$ (T3: rabbit).

²⁸The p -values based on WMW tests are: $p = 0.028$ (T1: control), $p = 0.958$ (T2: planner) and $p < 0.001$ (T3: rabbit).

²⁹Pairwise WMW tests yield the following p -values for nonprocrastinating students: $p < 0.001$ (T2: planner) and $p < 0.001$ (T3: rabbit). For procrastinating students, the results are: $p = 0.001$ (T2: planner) and $p < 0.001$ (T3: rabbit).

³⁰The p -values based on WMW tests are: $p = 0.580$ (T1: control), $p = 0.055$ (T2: planner) and $p = 0.046$ (T3: rabbit).

Figure 4.9: Actual Versus Planned Study Time on the MicroApp

(a) Condition II: Planner

(b) Condition III: Rabbit

Notes: The figure shows per experimental condition the daily actual (red triangle) and planned (green dot) study time on the MicroApp during the second stage of the study (when experimental variation was implemented), conditional on having logged in at least once during the week and having set a non-zero weekly study goal. The bars show the daily difference between actual and planned study time, with a negative value indicating that students on average fell short of their preset goal for that day. The hollow orange square shows the average daily actual study time of students in the control group (condition I). A graded quiz is administered during each practical class (P) on Friday.

in part planned beforehand. The overall impression from the figure is that students seem to have difficulties in sticking to their initial plans.

I explore this further in Figure 4.10. The figure shows the net balance of the daily differences between actual and planned study time. A negative end-of-week net balance indicates that students exerted lower effort by the end of the week than initially planned. I observe that in most cases the net balance is negative and growing in size during the week. This is in contrast to what one would expect to see when students catch up. In case of catching up, the negative net balance should become smaller (or turn positive) later on in the week (when the weekly study goal is due). The figure reveals that this is not the case among treated students. Furthermore, I not only observe that treated students have negative end-of-week net balances, but that these negative balances are also larger than in the control group. Earlier, I found no evidence that treated students study more or less on the MicroApp, which implies that the larger negative end-of-week net balances are to a large extent due to having set more ambitious study goals.

The previous figures provided some first visual evidence that treated students have difficulties in completing their weekly study goals. In Table 4.10, I contrast the performance of treated students with that of control students and identify the treatment effects on study goal completion rates. The table reports the logit marginal effects on the likelihood of achieving the study goal in a given week, conditional on having set a non-zero goal. The dependent variable is an indicator variable equal to one if student i achieved his or her study goal in week t , zero otherwise. The time period is from the fifth lecture week (start of the experimental period) until the exam week (end of the experimental period).

In column (1) of the table I find some initial, albeit weak, evidence that treated students are indeed *less* likely to reach their study goal by the end of the week, even after controlling for goal levels. In column (2), I restrict attention to those students who logged in once or more on the MicroApp during the week. These students have at least attempted to reach their goal, and, importantly, were exposed to the experimental variation on the MicroApp. Among these students, the negative effect persists and is larger in size: students in the planner condition and the rabbit condition are

Figure 4.10: Net Balance Actual Vs. Planned Study Time on the MicroApp

(a) Condition II: Planner

(b) Condition III: Rabbit

Notes: The dashed black line shows per experimental condition the net balance per day between the actual and planned study time on the MicroApp, conditional on having logged in at least once during the week and having set a non-zero weekly study goal. A negative value indicates that students are behind their own schedule for reaching the study goal. The study goal is set each week on Friday during an in-class survey and is due next week's Friday 23:59. The purple diamonds indicate the end-of-week net balance of students in the control group. Control students only submitted a weekly study goal, and not a daily study plan.

respectively about 13 and 14 percentage points less likely to reach their goal by the end of the week. When I do not control for the size of the study goal, as in column (3), these numbers change to 16 and 20 percentage points, respectively. This suggests that part of the negative effect is explained by treated students setting more ambitious study goals (as identified earlier). Similar results are obtained when the full sample of students is considered (that is, when no inclusion/exclusion rules are applied). I therefore find no support for hypothesis II in Section 4.3. To the contrary, I find some evidence that treated students are less likely to reach their study goals by the end of the week, partly as a result of setting more ambitious study goals than students in the control group.

Table 4.10: Treatment Effects on Study Goal Completion (Logit MEs)

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment=2	-0.045 (0.037)	-0.126** (0.059)	-0.155** (0.063)	-0.024 (0.030)	-0.095* (0.055)	-0.119** (0.058)
Treatment=3	-0.079** (0.038)	-0.139** (0.061)	-0.202*** (0.060)	-0.053* (0.031)	-0.110* (0.058)	-0.164*** (0.058)
Weekly study goal	-0.005 (0.003)	-0.026*** (0.005)		-0.002 (0.002)	-0.025*** (0.005)	
Number of students	283	211	211	403	251	251
Number of observations	875	432	432	1,194	496	496
Student inclusion/exclusion rules applied?	Yes	Yes	Yes	No	No	No
Conditional on weekly effort > 0	No	Yes	Yes	No	Yes	Yes

Notes: The table reports the treatment effects on the likelihood of achieving the study goal, conditional on having set a non-zero study goal. The time period under consideration is the experimental period (lecture week 5 - exam week). The dependent variable is an indicator variable equal to one if a student completed his or her study goal for the week, zero otherwise. Reported coefficients are logit marginal effects with robust standard errors in parentheses. The variable *study goal* is the study goal (in hours) set by student i in week t . ***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

In the appendix, I examine the treatment effects in more detail by exploring gender differences, differences between students identified as either high-performing or low-performing based on pre-experimental data, and differences between procrastinating students and nonprocrastinators. Figure 4.A.14 shows the gender-specific treatment effects. Previous studies have shown that goal-setting is more effective for males than for females (e.g., Clark et al. (2017)). While the point estimates in the figure suggest that the adverse treatment effects are indeed mainly observed among female students, the difference with male students is not significant.³¹ Figure 4.A.15 shows separate treatment effects for low-performing students and high-performing students. Among these two groups, I observe no dif-

³¹ Adding a gender dummy in the specification of column (2) in Table 4.10 and interacting this dummy with the treatment indicators yields the following p -values associated with the interaction effects: $p = 0.218$ (T2: planner) and $p = 0.618$ (T3: rabbit).

ferences in the treatment effects related to goal completion.³² Figure 4.A.16 distinguishes between procrastinators and nonprocrastinators. I find some evidence that nonprocrastinators in the planner group ($p = 0.037$) and the rabbit group ($p = 0.041$) are less likely to reach their goals compared to their counterparts in the control group. The treatment differences in goal completion rates between procrastinators and nonprocrastinators are not significantly different from zero.³³

Do the study goal completion rates of treated students improve over time during the experimental period? Each week, students observe their own goal progress and know by the end of the week whether they succeeded in completing the goal. In case of no success, students may be prompted to adjust their goal to more realistic levels. The act of formulating a study plan and observing one's own pacesetter may help in this process. Figure 4.11 shows the goal completion rates per week and experimental condition. Completion rates do improve over time, suggesting that students set more realistic goals. However, consistent with the previous results, the figure provides no support for the claim that these improvements are due to the treatment interventions, as the improvements are also observed among control students. In fact, even though all three conditions have similar low completion rates at the start of the experimental period, control students consistently achieve higher completion rates in subsequent weeks, giving rise to the negative treatment effects observed in Table 4.10.

Learning outcomes In Section 4.3 it was hypothesized that student effort serves as the mediator between the treatment interventions and learning outcomes. Insignificant treatment effects on student effort means that the supposed channel through which I expected improved learning outcomes is absent. This makes it unlikely to find support for the third hypothesis,

³² Adding a low-performance dummy in the specification of column (2) in Table 4.10 and interacting this dummy with the treatment indicators yields the following p -values associated with the interaction effects: $p = 0.696$ (T2: planner) and $p = 0.638$ (T3: rabbit).

³³ Adding a procrastination dummy in the specification of column (2) in Table 4.10 and interacting this dummy with the treatment indicators yields the following p -values associated with the interaction effects: $p = 0.294$ (T2: planner) and $p = 0.543$ (T3: rabbit).

Figure 4.11: Study Goal Completion Rates During the Experimental Period

(a) Cond. on study goal > 0 (b) Plus cond. on study time > 0

Notes: Study goal completion rates during the experimental period. Weekly completion rates are shown for each treatment condition separately. Figure (a) shows the rates for all students who submitted a non-zero study goal for the week. Figure (b) only considers those students who submitted a non-zero goal and additionally exerted non-zero study time on the MicroApp during the week.

which states that students who can observe their own pacesetter will have better learning outcomes than students who cannot, conditional on the pacesetter being successful in increasing student effort.

Figure 4.12 shows the learning outcomes per experimental condition. The final course grade is shown as well as the performance on each course grade component separately. A convenient feature of the research design is that I can compare learning outcomes before and after experimental variation is introduced. For instance, the midterm exam is administered before the start of the experiment and we should therefore not expect to see significant midterm performance differences between the experimental conditions. The figure shows that midterm scores are indeed balanced across the three groups.³⁴ The final exam is administered at the end of the experiment and focuses on course material covered during the experimental period. This makes final exam performance a suitable outcome variable for measuring the treatment effects on learning outcomes, as I also indicated in the pre-

³⁴Pairwise comparisons with the control group yield the following p -values (two-sided t -tests): $p = 0.606$ (T2: planner) and $p = 0.822$ (T3: rabbit).

analysis plan of this study. In the figure, I observe no evidence that treated students performed better on the final exam than control students.³⁵ This is in line with the previous reasoning that, absent increased student effort as the mediator, we should not expect to find a treatment effect on learning outcomes. Similar performance patterns are observed when looking at in-class quizzes and the final course grade of students.

The treatment effects on learning outcomes (final exam performance) are formally identified in Table 4.11. Looking at students who satisfy the inclusion rules of this study (columns (1) and (2)), I observe that the point estimates are small in size and not significantly different from zero. In analyses not shown for brevity I also looked at the final course grade as a measure of learning outcomes and found very similar results.

In the appendix, I present the treatment effects for each of the three subgroups discussed in this chapter. Table 4.A.5 shows the results separately for males and females. Female students scored on average about seven points higher on the final exam than male students. Interacting treatment- and gender indicators yields some (weak) evidence that the pacesetter feature triggered an adverse effect among female students. Compared to male students in the rabbit condition, female students who observed their own pacesetter scored an additional nine points lower on the final exam. Given that this result is only marginally significant (at the 10%-level), I view this as suggestive rather than conclusive evidence that the pacesetter caused the adverse effect among females. In Table 4.A.6 and Table 4.A.7, I respectively consider whether treatment effects differ for students based on their pre-experimental course performance or prior procrastination tendencies. The estimates show that this does not seem to be the case.

One concern with using increases in final exam scores as a proxy for improvements in learning outcomes is that it does not account for what students themselves wanted to achieve in the course. When the marginal cost of exerting extra effort exceeds the marginal benefit, a student may not find it worthwhile to go for a higher grade. In support of this, students indicated at the start of the course that they do not seek a perfect score on

³⁵Pairwise comparisons with the control group yield the following p -values (two-sided t -tests): $p = 0.812$ (T2: planner) and $p = 0.187$ (T3: rabbit).

Figure 4.12: Learning Outcomes per Experimental Condition

Notes: The figure shows the learning outcomes per experimental condition. For ease of comparison, all grades are converted to be on a 0-100 scale, with a higher score indicating better performance. The weight of each course grade component is as follows: midterm exam (30%), final exam (60%), and in-class quizzes (10%). The midterm exam (final exam) was administered before (after) the introduction of experimental variation. The final exam focused on material covered during the experimental period. Performance on in-class quizzes is the average of a student’s five best quizzes (out of six). Quizzes were administered every lecture week during the (pre-)experimental period. The error bars denote ± 1 standard error.

Table 4.11: Treatment Effects on Learning Outcomes (Final Exam Score)

	(1)	(2)	(3)	(4)
Treatment=2	-0.570 (2.390)	-0.475 (2.381)	-0.862 (1.845)	-0.797 (1.836)
Treatment=3	-3.045 (2.299)	-2.979 (2.306)	-4.198** (1.860)	-4.149** (1.859)
Constant	57.083*** (1.592)	56.248*** (1.820)	55.154*** (1.289)	54.060*** (1.385)
Number of students	302	302	504	504
Student inclusion/exclusion rules applied?	Yes	Yes	No	No
Control for second-stage effort on the MicroApp?	No	Yes	No	Yes

Notes: The table reports the treatment effects on learning outcomes. The dependent variable is the number of points scored on the final exam (out of 100 points). Second-stage effort on the MicroApp refers to the total study time of a student on the MicroApp during the experimental period. ***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

Table 4.12: Treatment Effects on Achieving Course Goals (Logit MEs)

	(1)	(2)	(3)
	Midterm Exam	Final Exam	Course Grade
Dependent variable:			
Treatment=2	-0.023 (0.066)	-0.015 (0.042)	-0.053 (0.047)
Treatment=3	-0.025 (0.065)	0.045 (0.046)	0.044 (0.053)
Number of students	309	309	309
Student inclusion/exclusion rules applied?	Yes	Yes	Yes

Notes: The table reports the treatment effects on the likelihood of achieving the initially formulated course goals. At the start of the course, each student was asked in an in-class survey to indicate their grade target for the midterm exam, the final exam, and the overall course. The dependent variable is equal to one if the actual grade is equal or higher than this target grade, zero otherwise. Reported coefficients are logit marginal effects. ***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

the exams but rather aim for a grade of around 7.5-8 (out of 10), as reported in Table 4.4. Therefore, in Table 4.12 I alternatively look at whether the treatments better enable students to reach their initial course goals. The dependent variable is equal to one if the actual grade is equal or higher than the initial target grade of the student, zero otherwise. I observe that the treatments do not increase the likelihood of achieving initial goals.

4.5.3 Additional analyses

Correlations between learning inputs and learning outputs One concern with targeting interventions at a particular learning input is that it may divert the attention of students from other, potentially more productive, learning inputs. In the previous analyses, I found no evidence that the treatments altered the way students studied in the course. No between-treatment differences were observed in terms of time investments in learning inputs. From a policy perspective, however, the question which learning inputs contribute to improved academic outcomes, and thus which inputs should be targeted, continues to be an important research topic that should inform the design and implementation of future educational interventions. With the collected data in this study, I can perform an exploratory analysis that serves as a first step in this direction.

The learning input that was targeted by the interventions in this study was usage of the MicroApp, an online learning platform with ample practice- and recap material. Motivating students to use a particular learning input

typically follows from the belief that the input improves learning outcomes. Since I did not randomize access to the MicroApp, I can not infer the causal impact the platform has had on student outcomes, and I can therefore not say to what extent this belief is warranted. What I can do with the available data is to look whether performance on the MicroApp correlates positively with performance in the course. Finding such a correlation is insightful for two reasons. One reason is that it would suggest that interim performance on the MicroApp has predictive power with regard to future learning outcomes, which can be of great value to researchers and educators who want to use educational technologies to identify, for example, at-risk students in real time. The second reason is that a positive correlation suggests that the relation between input and output goes in the expected direction.

In Figure 4.A.17, I plot the final course grade of students against their performance on the MicroApp. I restrict attention to those students who have solved at least 100 tasks on the app and measure performance as the percentage of correctly solved MicroApp practice tasks. The fitted regression line is upward sloping, indicating a positive correlation between performance in the course and on the MicroApp. The predicted values are intuitive: a student who has solved, say, 60% of MicroApp tasks correctly is expected to complete the course with a score of about six out of ten.

What is the productivity of each learning input in terms of improving learning outcomes? In Figure 4.A.18, I provide the results of an exploratory analysis that takes the following approach. For each lecture week, I use the in-class survey responses to determine for every student his or her time investment in a particular learning input. I then relate these time investments to the performance on in-class quizzes, covering the course material of that week. Specifically, I regress the quiz grade on student fixed effects and on each level of categorical variables that capture students' weekly time investment in a learning input. The reference category for an input consists of students with no time spent on that input during a specific week. For most inputs, there is no significant difference in quiz grades when more time is devoted to that input. Interestingly, I do find some evidence that quiz grades improve when more time is spent on the MicroApp, which suggests that the treatments were targeted at the right input.

4.6 Discussion

Technological innovations in education yield widespread excitement among policy makers. Research on these innovations, however, has not kept up with the pace at which these innovations are proliferating in educational programs, as noted by Escueta et al. (forthcoming). Most of the existing research has focused on comparing courses with digital learning components to more traditional courses with offline learning components. Swoboda and Feiler (2016), for example, find that student performance improves in courses that employ technological tools, such as online homework assignments and video lectures. Mixed results are reported in other studies.³⁶ More research is therefore needed that explores in detail how students interact with these technologies and how technologies should be designed to optimally improve the learning outcomes of students.

This chapter aimed to achieve just that. I developed a new educational technology to meticulously measure student effort, to contrast it with other learning inputs, and to administer personalized planning interventions to students in real time. I implemented a field experiment and used A/B-testing within the technology to identify the causal effects of different design features on student effort and outcomes. The empirical results shed light on how students interact with educational technologies, how their effort evolves throughout the course, and how outcomes are affected by technological and non-technological learning inputs. Subgroup analyses based on gender, ability and procrastination tendencies have further facilitated an understanding of these results.

The finding that the pacesetter feature does not alter effort and outcomes is highly relevant for the literature on commitment mechanisms in education. Most existing commitment mechanisms, such as deadlines and

³⁶Lee et al. (2010) find that students have favorable attitudes toward digital learning tools, and 73% of the students prefer these digital tools over traditional homework. Though, they find no evidence that online homework is more effective in improving test scores than traditional homework. Sosin et al. (2004) find a small but positive effect on student performance in classes that use technology. Hernandez-Julian and Peters (2012) find that students who submit homework online complete more assignments but have lower class attendance rates. They find no effect on exam scores when homework is completed online. Brown and Leidholm (2002) find that students in virtual classes perform significantly worse on exams than students in a live course.

goals, tend to be static and may not be based on students' own preferred study plans. The pacesetter, instead, dynamically links different selves by letting students "race" against their own initially preferred study pace. The pacesetter is thus fully consistent with a student's own initial plan. Furthermore, the pacesetter preserves flexibility, but still triggers the motivating effect of loss aversion when falling behind. While my analysis shows that students actively engage with their pacesetter, by setting non-zero study goals that are more ambitious than those set by control students, I find no evidence that this materializes into increased study effort. More research is needed that explores which type of pacesetter is most effective in bringing about changes in student effort and learning outcomes.

Through weekly surveys at the student level and MicroApp activity trackers, I have acquired detailed data on students' time investments in multiple learning inputs. This information is often limited or even absent in existing studies on student effort, forcing these studies to proxy effort by focusing on outcome measures instead. The data in this study yield rich insights on how students actually study. I observe that students invest most of their time in attending classes, reading offline course material, and practicing with offline material. With the exception of the MicroApp, students make surprisingly little use of other online resources. When exams are approaching, students shift most of their attention from "consuming" the course material to practicing. Comparing planned and actual study effort over time, I showed that students exert lower-than-planned effort right from the start and on average do not catch up later on.

In this study, I explored treatment heterogeneity in detail in order to see which specific groups benefit the most from the interventions. Duckworth et al. (2015) conjecture that educational interventions aimed at improving self-control may be especially beneficial for male students. I find no evidence of this for the specific interventions in this study. Male students in the treatment conditions do not study more, nor do they have improved learning outcomes. The absence of an effect is in contrast with Clark et al. (2017), who do find that task-based goals are more effective for male students than for female students. Considering treatment heterogeneity based on procrastination, Himmler et al. (2019) find that commitment devices

in education are most effective for procrastinating students. I find no evidence of this in the present field experiment. There is also no treatment heterogeneity when students are compared based on initial ability (pre-experimental performance in the course). Previous research has shown that incentives improve (worsen) the educational achievements of high-ability students (low-ability students) (Leuven, Oosterbeek, and van der Klaauw 2010). I find no treatment differences between low-ability students and high-ability students in a setting with non-financial incentives.

4.7 Conclusion

A lack of study time often prevents students from reaching their academic potential. While recent studies suggest that motivating students to study may be better achieved when effort is incentivized rather than output, only limited progress on this is made thus far due to a lack of data on effort and initial study plans. Procrastination tendencies further exacerbate the challenge of getting students to study more. These issues have stalled progress on designing and evaluating interventions that target student effort.

In a field experiment with 573 university students, I elicit detailed weekly study goals and plans from all students and use these to construct individualized pacesetters (*rabbits*). Pacesetters are moving reference points that visualize the preferred study pace of the present self by moving exactly according to the initial study plan. I build a new educational technology to measure student effort and to present the pacesetters to students in real time. Students can at any time see how they perform relative to their own pacesetter. Falling behind immediately makes salient to the student that he or she is reneging on earlier study commitments. The pacesetter thus serves as an internal commitment mechanism whereby future selves are triggered to follow the pacesetter and, in doing so, implement the initial study plan.

The experiment is designed to identify the causal effect of the pacesetter on student effort and learning outcomes, and to control for confounding effects related to the acts of formulating study goals and plans. The findings are as follows. First, I find that the pacesetter induces students to set more ambitious study goals, amounting to about two additional study hours per

week. This, however, does not lead to more effort exerted. I find that the pacesetter has no impact on study time, as measured by the educational technology, nor do I observe effort spillovers to other learning inputs and other courses. The richness of the data allows me to examine exactly when actual study time deviates from planned study time. I observe that students already fall behind their pacesetter at the start of the week and on average show no tendency to catch up later in the week. I consider multiple measures of learning outcomes and find no effect on these outcomes.

In additional analyses, I consider treatment heterogeneity based on gender, ability, and prior procrastination tendencies. In terms of student effort, I find no heterogeneity within these groups. With regard to goal-setting behavior and goal completion rates, I find some evidence that the pacesetter induces low-ability students (procrastinating students) to set higher (lower) weekly study goals. In all three groups, however, no differences are observed in the likelihood of actually achieving the weekly study goals. There are also no within-group differences observed for learning outcomes, with the exception of suggestive evidence that the pacesetter caused females to perform poorer on the exam.

This study is one of the first attempts to use educational technologies as a means to target behaviors that have a negative impact on academic performance. I show how these technologies can be used to measure student effort in real time, to administer personalized interventions, and to implement experimental evaluations through A/B-testing. In a review of the literature, Escueta et al. (forthcoming) conclude that technology-enabled interventions have great potential for improving education-related outcomes, but they also observe that existing studies have mostly limited their attention to campaigns with text messages. I show instead the potential of technologies to administer fully customized data-driven interventions at low marginal cost.

The present study points to several directions for future research. First, more studies are needed that explore how technology can be used to target behaviors that adversely affect academic performance. Such studies will be particularly fruitful when they utilize the unique benefits that technology has to offer, from meticulous measurement of student behavior to

implementing personalized interventions. While this study has focused on procrastination and self-control problems among students, I reckon that other behaviors are suitable for further investigation as well. Second, this study has used students' own initial study plans as input for the construction of the pacesetter. Interesting possibilities for future research are to use alternative sources that determine the pace of the pacesetter.

The digitalization of education unveils many new opportunities and challenges for researchers and educators alike. The ease and frequency with which students use technology in their everyday lives likely results in that studying will increasingly take place online. How students learn in a digital world is an important question with profound implications for how human capital will be accumulated in the future.

4.A Appendix

4.A.1 Screenshots of the MicroApp (pre-experiment)

Figure 4.A.1: Dashboard With Basic Learning Analytics

Figure 4.A.2: Answered Practice Task (With Feedback)

Figure 4.A.3: Recap on the Course Topic of Taxes and Subsidies

4.A.2 A sample in-class survey (baseline period)

Figure 4.A.4: Frontside of the Survey

Microeconomics for E&BE: Consumers & Firms
Weekly Survey

- Your answers to the survey questions remain private and will be handled confidentially.
- Your answers will be used solely for improving the course structure and for general research purposes.
- Please answer truthfully. Your answers will not in any way affect your final course grade.
- For questions about the survey, please contact the secretary: Kimberley Vudinh (k.m.vudinh@rug.nl).

This survey is about your study behavior in the past seven days (Saturday 15 September – Friday 21 September)

Q1: For this course, Microeconomics for E&BE, please indicate your approximate time investment in the past seven days for each of the following categories (where < means “less than”):

Category	Time investment
Attending classes (lecture, tutorial, practical)	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours
Reading offline course material (e.g., textbook, lecture slides, hand-outs)	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours
Practicing with offline course material (e.g., tutorial exercises, practice exams, etc.)	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours
Use of the MicroApp	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours
Reading/watching online course material (not Microapp!) (e.g., YouTube tutorial videos)	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours
Practicing with online course material (not Microapp!) (e.g., websites with practice exercises)	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours
Any study activity that is not captured by one of the categories mentioned above (e.g., attending an office hour of a lecturer)	☐ Not at all ☐ Less than 1 hour ☐ 1 - <3 hour(s) ☐ 3 - <5 hours ☐ 5 hours or more than 5 hours

Figure 4.A.5: Backside of the Survey

Q2: Please answer the following if you take other courses this block. With the categories from the previous question in mind, what was your approximate average time investment for these other courses (note again that < means "less than")?

- ☐ Not at all
- ☐ Less than 5 hours
- ☐ 5 - <10 hours
- ☐ 10 - <15 hours
- ☐ 15 hours or more than 15 hours

Example: let's say you follow two other courses: course A and course B. If your time investment last week was $\bar{X}$ hours for course A and $\bar{Y}$ hours for course B, your average time investment for these courses was: $\frac{\bar{X} + \bar{Y}}{2}$.

Thank you for filling in the survey!

- End of the survey -

4.A.3 Experimental variation in the in-class surveys

Figure 4.A.6: Experimental Variation in the In-Class Surveys

Your MicroApp study goal [All experimental conditions]

Q3: Challenge yourself by setting a goal for how many hours you intend to make use of the MicroApp during your self-study in the upcoming week. Fill in the blanks below. Note that half hours can be written as 0.5.

In the upcoming seven days, my goal is to practice for hours on the MicroApp before Friday 19 October 23:59.

- From tomorrow on, you can track your progress toward your study goal via www.microapp.nl.

Your preferred MicroApp study schedule [Conditions II and III]

Q4: In order to reach your MicroApp study goal, please indicate for each day in the table below your daily goal in hours, and how you intend to allocate these hours over the day. Make sure that your daily goals add up to your study goal stated in Q3. Note again that half hours can be written as 0.5.

	Saturday 13 October	Sunday 14 October	Monday 15 October	Tuesday 16 October	Wednesday 17 October	Thursday 18 October	Friday 19 October
Daily goal in hours							
Night 00:00-06:00							
Morning 06:00-12:00							
Afternoon 12:00-18:00							
Evening 18:00-00:00							

- From tomorrow on, you can consult your preferred study schedule at any time via www.microapp.nl. [Conditions II and III]
- The MicroApp now contains a *pacesetter* feature. The pacesetter is a moving point that implements your study goal by moving exactly according to your own preferred study schedule. From tomorrow on, you can see at any time where you stand relative to your own pacesetter via www.microapp.nl. [Condition III]

4.A.4 Questions on course goals in the first in-class survey

Figure 4.A.7: Survey Questions on Course Goals

Q3: Please answer the following sub-questions related to your goals in this course (Microeconomics for E&BE). Note that the Dutch grading system is based on a ten-point scale, where 1 means very poor and 10 means excellent.

Q3.1: Which overall grade are you aiming for in this course?

Q3.2: Which grade are you aiming for on the mid-term exam (Tuesday 2 October, 2018)?

Q3.3: Which grade are you aiming for on the final exam (Thursday 1 November, 2018)?

Q3.4: What are the chances of you participating in the final exam [0-100%]?

4.A.5 Intertemporal choices about real effort in the third in-class survey

Figure 4.A.8: Intertemporal Choice Task About Real Effort

A research team at this faculty is currently recruiting students for completing an online survey. Students who complete the survey have **a chance of winning one of the 10 available €50 Bol.com gift certificates**. Bol.com (www.bol.com) is the leading web shop in the Netherlands for books, toys and electronics. The available gift certificates are **exclusively for students in this course!**

Please indicate your preferences for filling in the survey by **making your decision in each of the 10 choices below**. One of your choices may be the survey you will have to answer to have a chance of winning a €50 Bol.com gift certificate. Indicating your preferences below does not mean that you are required to complete the online survey (send to you via email). The date of receiving the gift certificate does not depend on when you finish the survey (as long as you completed before the deadline).

The survey can either be answered in the **next 5 days (short survey)** or in the **next 35 days (long survey)**

Short version of the online survey Before Thursday, 4 October 2018			Long version of the online survey Before Saturday, 3 November 2018		
<i>Length of the survey</i>			<i>Length of the survey</i>		
Choice #1	☐	15 minutes	or	☐	30 minutes
Choice #2	☐	18 minutes	or	☐	30 minutes
Choice #3	☐	21 minutes	or	☐	30 minutes
Choice #4	☐	24 minutes	or	☐	30 minutes
Choice #5	☐	27 minutes	or	☐	30 minutes

The survey can either be answered in the **next 90 days (short survey)** or in the **next 120 days (long survey)**

Short version of the online survey Before Friday, 28 December 2018			Long version of the online survey Before Sunday, 27 January 2019		
<i>Length of the survey</i>			<i>Length of the survey</i>		
Choice #6	☐	15 minutes	or	☐	30 minutes
Choice #7	☐	18 minutes	or	☐	30 minutes
Choice #8	☐	21 minutes	or	☐	30 minutes
Choice #9	☐	24 minutes	or	☐	30 minutes
Choice #10	☐	27 minutes	or	☐	30 minutes

4.A.6 Student enrollment dates in the course

Figure 4.A.9: Student Enrollment Dates in the Course

Notes: The figure shows the student enrollment dates in the course. Students who enrolled before the registration deadline (26 August 2018) are guaranteed a spot in the course. Enrolling after the deadline is also possible, but it is not guaranteed that a spot is available. In practice, this policy is not too strict and students can generally still take the course when they enroll after the deadline.

4.A.7 MicroApp study time: intention-to-treat (ITT) estimates

Table 4.A.1: Treatment Effects on MicroApp Study Time (Intention-to-Treat Estimates)

	(1)	(2)	(3)	(4)	(5)
Post=1	51.279*** (9.549)	51.279*** (9.553)	37.615*** (6.181)	37.615*** (6.183)	0.678*** (0.151)
Post=1 $\times$ Treatment=2	-24.255* (13.663)	-24.255* (13.668)	-5.729 (9.029)	-5.729 (9.030)	-0.435* (0.239)
Post=1 $\times$ Treatment=3	-18.099 (13.521)	-18.099 (13.526)	-16.258* (8.697)	-16.258* (8.699)	-0.364 (0.238)
Treatment=2		-3.268 (13.630)		-4.141 (8.685)	
Treatment=3		3.003 (13.357)		2.791 (8.831)	
Constant	90.588*** (2.780)	90.605*** (9.645)	52.208*** (1.817)	52.642*** (6.291)	4.265*** (0.049)
R ²	.00866	.0088	.00758	.00783	.025
Number of students	309	309	569	569	305
Number of observations	2,472	2,472	4,552	4,552	1,174
Student fixed effects	Yes	No	Yes	No	Yes
Student inclusion/exclusion rules applied?	Yes	Yes	No	No	Yes
Time unit	Week	Week	Week	Week	Week
Dependent variable	Level	Level	Level	Level	Log

Notes: Identification of the intention-to-treat treatment effects on MicroApp study time (in minutes). Robust standard errors (clustered at the student level) are in parentheses. The variable *post* captures the experimental period. Study time during the transition day (5 October 2018) from the baseline period (lecture weeks 1-4) to the experimental period (lecture weeks 5-8) is excluded. There are two exclusion rules. First, a student is excluded when he or she participated in less than two practicals after the midterm exam. Second, a student is also excluded when he or she never made use of the MicroApp in the period until the first practical after the midterm exam. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

4.A.8 MicroApp study time: males versus females

Table 4.A.2: Treatment Effects on MicroApp Study Time, by Gender

	(1)	(2)	(3)	(4)	(5)
Post=1	123.365*** (32.740)	138.784*** (31.150)	112.413*** (30.764)	130.489*** (28.530)	0.749*** (0.203)
Post=1 $\times$ Treatment=2	-82.904** (41.698)	-92.817** (39.346)	-65.418* (39.514)	-75.307** (36.888)	-0.586* (0.322)
Post=1 $\times$ Treatment=3	-72.218* (41.326)	-69.889* (38.876)	-55.656 (39.742)	-56.767 (36.328)	-0.342 (0.267)
Post=1 $\times$ Gender=1	-59.329 (44.293)	-53.658 (40.557)	-40.256 (42.772)	-58.137 (38.607)	-0.304 (0.303)
Post=1 $\times$ Treatment=2 $\times$ Gender=1	71.775 (60.823)	74.742 (53.892)	50.428 (58.106)	91.183* (51.742)	0.568 (0.445)
Post=1 $\times$ Treatment=3 $\times$ Gender=1	102.298 (75.706)	144.450** (73.638)	86.432 (73.276)	116.109* (68.731)	-0.056 (0.687)
Treatment=2		12.423 (30.014)		17.396 (27.421)	
Treatment=3		43.103 (29.115)		36.302 (26.518)	
Gender=1		18.225 (31.965)		11.925 (29.802)	
Treatment=2 $\times$ Gender=1		-40.660 (45.304)		-45.887 (42.203)	
Treatment=3 $\times$ Gender=1		-73.507 (47.159)		-51.724 (44.142)	
Constant	196.998*** (6.303)	174.118*** (20.401)	186.983*** (6.355)	165.859*** (18.525)	4.270*** (0.049)
R ²	.0307	.0373	.0291	.0333	.0262
Number of students	300	300	404	404	300
Number of observations	1,148	1,148	1,343	1,343	1,148
Student fixed effects	Yes	No	Yes	No	Yes
Student inclusion/exclusion rules applied?	Yes	Yes	No	No	Yes
Time unit	Week	Week	Week	Week	Week
Dependent variable	Level	Level	Level	Level	Log

Notes: Identification of the gender-specific treatment effects on MicroApp study time (in minutes). The gender dummy variable is equal to one for female students, zero otherwise. Robust standard errors (clustered at the student level) are in parentheses. The variable *post* captures the experimental period. Study time during the transition day (5 October 2018) from the baseline period (lecture weeks 1-4) to the experimental period (lecture weeks 5-8) is excluded. There are two exclusion rules. First, a student is excluded when he or she participated in less than two practicals after the midterm exam. Second, a student is also excluded when he or she never made use of the MicroApp in the period until the first practical after the midterm exam. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

4.A.9 MicroApp study time: high-performance students versus low-performance students

Table 4.A.3: Treatment Effects on MicroApp Study Time, by Pre-Experimental Performance

	(1)	(2)	(3)	(4)	(5)
Post=1	87.907*** (27.085)	102.566*** (24.748)	75.846*** (26.052)	81.445*** (23.421)	0.595*** (0.186)
Post=1 $\times$ Treatment=2	-77.461** (36.758)	-70.832** (33.434)	-53.901 (35.800)	-39.209 (32.043)	-0.631** (0.257)
Post=1 $\times$ Treatment=3	-37.703 (38.824)	-13.136 (33.827)	-26.184 (37.687)	3.921 (32.708)	-0.540** (0.271)
Post=1 $\times$ Low performance=1	41.061 (48.015)	40.109 (44.143)	58.778 (45.336)	65.040 (40.481)	0.188 (0.311)
Post=1 $\times$ Treatment=2 $\times$ Low performance=1	50.160 (62.607)	13.251 (58.321)	16.312 (59.175)	-11.579 (53.961)	0.546 (0.514)
Post=1 $\times$ Treatment=3 $\times$ Low performance=1	-18.220 (67.866)	-46.720 (64.721)	-21.016 (65.449)	-70.685 (59.222)	0.453 (0.501)
Treatment=2		3.946 (31.646)		4.358 (29.465)	
Treatment=3		2.903 (29.889)		-0.427 (28.111)	
Low performance=1		-40.877 (31.145)		-52.316* (28.543)	
Treatment=2 $\times$ Low performance=1		-15.155 (43.406)		-2.633 (40.572)	
Treatment=3 $\times$ Low performance=1		45.913 (46.252)		52.178 (42.375)	
Constant	195.454*** (6.234)	198.094*** (19.212)	184.573*** (6.359)	192.723*** (18.191)	4.266*** (0.048)
R ²	.0236	.0366	.0214	.0331	.0125
Number of students	305	305	409	409	305
Number of observations	1,174	1,174	1,372	1,372	1,174
Student fixed effects	Yes	No	Yes	No	Yes
Student inclusion/exclusion rules applied?	Yes	Yes	No	No	Yes
Time unit	Week	Week	Week	Week	Week
Dependent variable	Level	Level	Level	Level	Log

Notes: Identification of the treatment effects on MicroApp study time (in minutes), conditional on the pre-experimental performance of students in the course. Robust standard errors (clustered at the student level) are in parentheses. Pre-experimental course performance is measured as the within-treatment relative performance of a student on the midterm exam. The bottom (top) 50% is the low-performing (high-performing) group. Only students that satisfy the inclusion rules of this study are considered when calculating relative performances. The low-performance dummy variable is equal to one when a student is in the within-treatment bottom 50% in terms of pre-experimental performance, zero otherwise. The variable *post* captures the experimental period. Study time during the transition day (5 October 2018) from the baseline period (lecture weeks 1-4) to the experimental period (lecture weeks 5-8) is excluded. ** $p < 0.01$, * $p < 0.05$, $p < 0.1$

4.A.10 MicroApp study time: procrastinators versus non-procrastinators

Table 4.A.4: Treatment Effects on MicroApp Study Time, by Procrastination

	(1)	(2)	(3)	(4)	(5)
Post=1	94.569*** (30.493)	115.497*** (28.869)	94.206*** (29.632)	100.667*** (27.105)	0.651*** (0.180)
Post=1 × Treatment=2	-67.784 (41.933)	-71.003* (38.544)	-63.467 (40.293)	-42.742 (37.015)	-0.651** (0.306)
Post=1 × Treatment=3	-64.047 (40.473)	-53.109 (38.237)	-63.973 (39.645)	-45.828 (36.437)	-0.415 (0.290)
Post=1 × Procrastinator=1	41.606 (44.241)	13.879 (40.367)	28.265 (42.593)	27.051 (38.558)	0.133 (0.338)
Post=1 × Treatment=2 × Procrastinator=1	4.121 (59.140)	11.884 (53.566)	22.089 (56.669)	-1.653 (51.379)	0.472 (0.498)
Post=1 × Treatment=3 × Procrastinator=1	38.448 (65.930)	50.474 (60.896)	62.671 (63.961)	45.302 (57.237)	0.082 (0.514)
Treatment=2		7.797 (31.621)		3.849 (29.506)	
Treatment=3		32.417 (29.953)		32.080 (28.946)	
Procrastinator=1		-41.326 (29.011)		-49.517* (26.424)	
Treatment=2 × Procrastinator=1		-23.501 (41.780)		-2.359 (38.822)	
Treatment=3 × Procrastinator=1		-26.660 (44.233)		-23.588 (40.380)	
Constant	195.932*** (6.224)	196.688*** (20.255)	184.252*** (6.268)	189.188*** (19.425)	4.266*** (0.049)
R ²	.0212	.0408	.0188	.0363	.0185
Number of students	303	303	410	410	303
Number of observations	1,167	1,167	1,365	1,365	1,167
Student fixed effects	Yes	No	Yes	No	Yes
Student inclusion/exclusion rules applied?	Yes	Yes	No	No	Yes
Time unit	Week	Week	Week	Week	Week
Dependent variable	Level	Level	Level	Level	Log

Notes: Identification of the treatment effects on MicroApp study time (in minutes), conditional on prior procrastination tendencies in the course. A student is identified as a procrastinator when his or her course enrollment date is after the within-treatment median course enrollment date. Robust standard errors (clustered at the student level) are in parentheses. Only students that satisfy the inclusion rules of this study are considered when calculating relative performances. The procrastinator dummy variable is equal to one when a student enrolled in the course after the within-treatment median course enrollment date, zero otherwise. The variable *post* captures the experimental period. Study time during the transition day (5 October 2018) from the baseline period (lecture weeks 1-4) to the experimental period (lecture weeks 5-8) is excluded.

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

4.A.11 Second-stage time investments in learning inputs

Figure 4.A.10: Second-Stage Time Investments in Learning Inputs

Notes: Students' time investments in learning inputs are elicited at the end of each lecture week via in-class surveys. In the second stage of the experiment (overlapping with the last three lecture weeks and an exam week) there is experimental variation in the MicroApp and the in-class surveys (see Section 4.4.3). The final exam (worth 60% of the final grade) was in the exam week (the week after the seventh lecture week). No in-class survey was administered during the exam week. The learning input "classes" is partly incentivized (attendance is required to be eligible for a participation grade (worth 10% of the final grade)). The MicroApp is a newly introduced online learning platform in the course with abundant practice- and recap material. The categories "read/watch online" and "practice online" refer to non-MicroApp online study activities.

4.A.12 Goal-setting: males versus females

Figure 4.A.11: Weekly Study Goals for MicroApp Usage, by Gender

Notes: The figure shows per experimental condition how many hours students on average plan to study on the MicroApp in each week during the experimental period, conditional on having set a non-zero study goal. Study goals are elicited through weekly in-class surveys. The average of these study goals is reported in the figure, split by gender and treatment status. The error bars denote ± 1 standard error.

4.A.13 Goal-setting: high-performance students versus low-performance students

Figure 4.A.12: Weekly Study Goals for MicroApp Usage, by Prior Performance

Notes: The figure shows per experimental condition how many hours students on average plan to study on the MicroApp in each week during the experimental period, conditional on having set a non-zero study goal. Study goals are elicited through weekly in-class surveys. The average of these study goals is reported in the figure, split by treatment status and students' pre-experimental performance in the course. Pre-experimental course performance is measured as the within-treatment relative performance of a student on the midterm exam. The bottom (top) 50% is the low-performing (high-performing) group. The error bars denote ± 1 standard error.

4.A.14 Goal-setting: procrastinators versus nonprocrastinators

Figure 4.A.13: Weekly Study Goals for MicroApp Usage, by Prior Procrastination

Notes: The figure shows per experimental condition how many hours students on average plan to study on the MicroApp in each week during the experimental period, conditional on having set a non-zero study goal. Study goals are elicited through weekly in-class surveys. The average of these study goals is reported in the figure, split by treatment status and students’ prior procrastination tendencies in the course. The within-treatment median course enrollment date is used as the cutoff for identifying procrastinators (applied after) and nonprocrastinators (applied before) in each treatment. The error bars denote ± 1 standard error.

4.A.15 Goal completion: males versus females

Figure 4.A.14: Treatment Effects on Study Goal Completion (logit MEs), by Gender

(a) Conditional on weekly study goal > 0

(b) Plus conditional on weekly MicroApp study time > 0

Notes: The figure shows the gender-specific treatment effects on the likelihood of achieving the study goal, after controlling for the goal level. The time period under consideration is the experimental period (lecture week 5 - exam week). The dependent variable is an indicator variable equal to one if a student completed his or her study goal for the week, zero otherwise. The coefficients shown are logit marginal effects. The reference group consists of students in the control group (T1). The vertical spikes indicate 95% confidence intervals.

4.A.16 Goal completion: high-performance students versus low-performance students

Figure 4.A.15: Treatment Effects on Study Goal Completion (logit MEs), by Prior Performance

(a) Conditional on weekly study goal > 0

(b) Plus conditional on weekly MicroApp study time > 0

Notes: The figure shows the treatment effects on the likelihood of achieving the study goal, after controlling for the goal level. Treatment effects are shown separately for students with high/low pre-experimental performance. Low (high) performers are students in the within-treatment bottom (top) 50% based on performance on the midterm exam. The time period under consideration is the experimental period (lecture week 5 - exam week). The dependent variable is an indicator variable equal to one if a student completed his or her study goal for the week, zero otherwise. The coefficients shown are logit marginal effects. The reference group consists of students in the control group (T1). The vertical spikes indicate 95% confidence intervals.

4.A.17 Goal completion: procrastinators versus nonprocrastinators

Figure 4.A.16: Treatment Effects on Study Goal Completion (logit MEs), by Procrastination

(a) Conditional on weekly study goal > 0

(b) Plus conditional on weekly MicroApp study time > 0

Notes: The figure shows the treatment effects on the likelihood of achieving the study goal, after controlling for the goal level. Treatment effects are shown separately for students with and without prior procrastination tendencies. I identify procrastination tendencies based on whether students enrolled for the course before or after the within-treatment median course enrollment date. The time period under consideration is the experimental period (lecture week 5 - exam week). The dependent variable is an indicator variable equal to one if a student completed his or her study goal for the week, zero otherwise. The coefficients shown are logit marginal effects. The reference group consists of students in the control group (T1). The vertical spikes indicate 95% confidence intervals.

4.A.18 Treatment effects on final exam performance: males versus females

Table 4.A.5: Treatment Effects on Learning Outcomes (Final Exam Performance), by Gender

	(1)	(2)	(3)	(4)
Treatment=2	-0.457 (2.887)	-0.411 (2.885)	-1.250 (2.184)	-1.217 (2.186)
Treatment=3	-0.341 (2.742)	-0.332 (2.762)	-2.970 (2.139)	-2.961 (2.151)
Gender=1	7.035** (3.477)	6.983** (3.486)	3.470 (2.988)	3.368 (2.976)
Treatment=2 $\times$ Gender=1	0.404 (5.038)	0.550 (5.041)	1.714 (4.114)	1.821 (4.088)
Treatment=3 $\times$ Gender=1	-9.067* (5.100)	-8.853* (5.137)	-4.170 (4.473)	-3.950 (4.473)
Constant	54.935*** (1.895)	54.243*** (2.058)	54.167*** (1.497)	53.143*** (1.567)
Number of students	297	297	491	491
Student inclusion/exclusion rules applied?	Yes	Yes	No	No
Control for second-stage effort on the MicroApp?	No	Yes	No	Yes

Notes: The table reports the gender-specific treatment effects on learning outcomes. The dependent variable is the number of points scored on the final exam (out of 100 points). The indicator variable *gender* equals one for female students, zero otherwise. ***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

4.A.19 Treatment effects on final exam performance: high-performance students versus low-performance students

Table 4.A.6: Treatment Effects on Learning Outcomes (Final Exam Performance), by Prior Performance

	(1)	(2)	(3)	(4)
Treatment=2	1.296 (2.418)	1.297 (2.417)	1.083 (1.983)	1.095 (1.971)
Treatment=3	-3.765 (2.461)	-3.728 (2.468)	-4.412** (2.118)	-4.436** (2.118)
Low performance=1	-18.581*** (2.617)	-18.577*** (2.626)	-18.244*** (2.144)	-18.171*** (2.148)
Treatment=2 $\times$ Low performance=1	-4.415 (3.851)	-4.313 (3.848)	-4.625 (2.983)	-4.572 (2.969)
Treatment=3 $\times$ Low performance=1	-0.804 (3.893)	-0.821 (3.901)	0.186 (3.138)	0.285 (3.144)
Constant	65.073*** (1.754)	64.678*** (1.943)	63.898*** (1.450)	63.085*** (1.529)
Number of students	302	302	508	508
Student inclusion/exclusion rules applied?	Yes	Yes	No	No
Control for second-stage effort on the MicroApp?	No	Yes	No	Yes

Notes: The table reports the treatment effects on learning outcomes. The dependent variable is the number of points scored on the final exam (out of 100 points). Treatment effects are shown separately for students with high/low pre-experimental performance (as measured by the grade on the midterm exam). Low (high) performers are students in the within-treatment bottom (top) 50% based on midterm scores. ***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

4.A.20 Treatment effects on final exam performance: procrastinators versus nonprocrastinators

Table 4.A.7: Treatment Effects on Learning Outcomes (Final Exam Performance), by Procrastination

	(1)	(2)	(3)	(4)
Treatment=2	0.253 (3.117)	0.373 (3.114)	-0.181 (2.648)	0.033 (2.637)
Treatment=3	-2.742 (2.926)	-2.726 (2.929)	-2.242 (2.488)	-2.137 (2.491)
Procrastinator=1	-2.542 (3.242)	-2.381 (3.311)	-3.234 (2.516)	-2.673 (2.567)
Treatment=2 $\times$ Procrastinator=1	-2.279 (4.825)	-2.411 (4.844)	-1.476 (3.645)	-1.857 (3.652)
Treatment=3 $\times$ Procrastinator=1	-0.692 (4.697)	-0.634 (4.725)	-3.718 (3.653)	-3.911 (3.664)
Constant	58.179*** (2.042)	57.488*** (2.358)	56.567*** (1.843)	55.399*** (1.992)
Number of students	301	301	508	508
Student inclusion/exclusion rules applied?	Yes	Yes	No	No
Control for second-stage effort on the MicroApp?	No	Yes	No	Yes

Notes: The table reports the treatment effects on learning outcomes. The dependent variable is the number of points scored on the final exam (out of 100 points). Treatment effects are shown separately for students with and without prior procrastination tendencies. I identify procrastination tendencies based on whether students enrolled for the course before or after the within-treatment median course enrollment date. ***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

4.A.21 Correlation between student performance in the course and on the MicroApp

Figure 4.A.17: Student Performance in the Course and on the MicroApp

Notes: The figure shows a scatterplot of student performance in the course (unrounded final course grade) and on the MicroApp (percentage of correctly solved practice tasks). The course grade is based on performance on a midterm exam (30%), a final exam (60%), and weekly in-class quizzes (10%). A higher grade indicates better performance. Only students who have solved at least 100 MicroApp practice tasks are included in the figure.

4.A.22 Time investments in learning inputs and performance on weekly quizzes

Figure 4.A.18: Performance on Weekly Quizzes, by Time Investments in Learning Inputs

Notes: The figure plots the coefficients of a regression where the weekly in-class quiz grade (from 0 (poor) to 10 (excellent)) of a student is regressed on student fixed effects and on each level of categorical variables that capture students' weekly time investment in each learning input. The horizontal spikes indicate 95% confidence intervals. Standard errors are clustered at the student level. The reference category for each learning input consists of students with no time investment in that learning input during a specific week. The time investments of each student are elicited through end-of-week in-class surveys. Quizzes are administered at the end of each lecture week (Friday) and cover the course material of that week.

Chapter 5

P(l)aying at the pump: Incentivized measurement of risk attitudes in the field*

Abstract

Decisions that involve uncertainty are shaped by individual risk attitudes. Valid measures of risk attitudes are thus crucial for research and policy. In the field, these measures are typically elicited through non-incentivized surveys because time and complexity hinder the use of incentive-compatible experimental methods. Do these two approaches yield similar assessments of risk attitudes in a natural field setting with a large group of consumers? Are the estimated risk attitudes stable in repeat elicitations? We collaborate with a startup in the retail gasoline industry and for one year use their on-site pump displays to repeatedly administer both incentivized (Bomb Risk Elicitation Task) and non-incentivized (SOEP risk question) risk elicitation methods to about 2,500 gas station visitors. We find that more than 80% of the subjects make decisions in the BRET that are consistent with risk aversion. No gender differences are observed in the BRET, nor do other demographics explain some of the variance. The correlation with responses to the SOEP question is positive, but small in magnitude. In repeat elicitations, we observe time-varying individual risk attitudes.

*This chapter is based on Romensen and Soetevent (2020b).

5.1 Introduction

Given the central role played by individual risk attitudes in many decisions involving uncertainty, it is not surprising that by now an extensive literature has emerged that examines how risk attitudes can be empirically measured. A key challenge identified in this literature is that there seems to be a trade-off between the practical feasibility of eliciting the risk attitudes of large groups in the field and making sure that the elicited attitudes are incentive compatible (Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner 2011). Two separate approaches have therefore become popular, each emphasizing a different side of the trade-off. Survey-based approaches are fast and relatively inexpensive, but generally not incentive compatible. Approaches based on experimental methods are more precise and incentive compatible, but they are rarely used in field studies due to time and cost considerations, complexity, and infrastructure needs.¹ Do the two approaches yield similar qualitative estimates of individual risk attitudes in a natural field setting?

In this chapter, we show that the adoption of digital technologies in business-to-consumer (B2C) communication increases the scope for the incentive-compatible measurement of individual risk attitudes in the field. We collaborate with a startup in the retail gasoline industry that has installed pump displays with touch screens at its filling stations.² With these displays, the company can communicate with its customers in real time and store all responses. We use this on-site infrastructure for one year to repeatedly administer two risk elicitation methods among a large and diverse group of consumers: the Bomb Risk Elicitation Task and the SOEP survey question on willingness to take risk.

The Bomb Risk Elicitation Task (BRET) is an experimental method. In the BRET, introduced and validated by Crosetto and Filippin (2013),

¹For example, Charness, Gneezy and Imas (2013, p. 44) comment on the Balloon Analogue Risk Task (BART), a popular experimental risk elicitation method, that "...it may not be ideal for being paired with other decision-making tasks, when time is a factor, or for field work where access to computers is limited."

²A filling station is arguably one of the best natural settings to find a wide cross-section of the total population of those aged above the legal driving age. In the Netherlands, as of 2018 there are 487 vehicles per 1,000 individuals. Most of these vehicles run on gasoline (79%) or diesel (16%) and thus require regular fill-ups. Data are retrieved from Statistics Netherlands (CBS).

subjects decide how many boxes to collect out of a set of 100, one of which contains a bomb. Subject earnings are linearly increasing with the number of boxes collected, but are zero if the box with the bomb is among the collected boxes. The number of boxes a subject collects provides a measure of individual risk attitudes. Advantages of the BRET are that minimal numerical skills are required, both risk aversion and risk seeking can be precisely estimated, and task measurements are unaffected by loss aversion due to the absence of a reference point other than zero; the task is entirely defined in the gain domain. For these reasons, the BRET compares favorably to many other commonly used experimental elicitation tasks. The task is popular in laboratory studies, and has recently been used in the context of research on, for example, peer effects (Gioia 2017), school choice (Basteck and Mantovani 2018), and cooperation (Hillenbrand and Winter 2018).

The SOEP question is part of the German Socio-Economic Panel Study (SOEP) and is a popular non-incentivized method to elicit risk attitudes in a timely and cost-effective manner. The exact wording of the question is as follows (from Dohmen et al. (2011)): “*How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?*”. Subjects can answer on a scale from 0 (not at all willing to take risks) to 10 (very willing to take risks). The SOEP question has been used in the field to study, among other topics, the relation between risk attitudes and migration (Jaeger, Dohmen, Falk, Huffman, Sunde, and Bonin 2010), self-employment (Skriabikova, Dohmen, and Kriechel 2014), and entrepreneurship (Koudstaal, Sloof, and van Praag 2016).

This chapter examines whether an incentive-compatible risk elicitation method (BRET) yields similar assessments of individual risk attitudes as the non-incentivized SOEP risk question in a natural field setting with a large non-student sample. Weimann and Brosig-Koch (2019) review the literature in experimental economics and observe that there are differences in behavior between students and non-students. They conclude that “...the question of transferability of observations made with students to the general population has not yet been settled conclusively” (2019, p. 65). Our findings contribute to this ongoing discussion.

About 2,500 visitors of thirteen geographically dispersed filling stations in the Netherlands completed the two elicitation methods on the pump displays. The main findings are as follows. First, the vast majority of subjects, about 82%, submit choices in the BRET that are consistent with risk aversion, which is about 30 percentage points higher than the share reported by Crosetto and Filippin (2013) in their lab studies with a student sample.³ We observe no gender differences in BRET choices, nor do we observe that other demographic factors, such as age and proxies for wealth, can explain some of the variance in the number of boxes collected in the BRET. Second, the modal response to the SOEP risk question is seven, which suggests that, according to their own self-assessment, subjects appear willing to take risks. This is particularly so for male subjects. The correlation between choices in the BRET and responses to the SOEP question is found to be positive but small. Finally, in repeated rounds of both the BRET and the SOEP question, subjects appear to become more willing to take risks, which is suggestive of time-varying individual risk attitudes.

In traditional economic theory, it is commonly assumed that individual risk attitudes are an innate trait that can be captured by a single time-invariant parameter.⁴ A growing body of empirical evidence, however, suggests that risk attitudes are time-varying (Chuang and Schechter 2015) and domain-specific (Einav, Finkelstein, Pascu, and Cullen 2012). The correlation of an individual's risk attitudes across time is positive but typically moderate (Schildberg-Hörisch 2018). In light of this evidence, having access to just-in-time estimates of individual risk attitudes in the direct decision-making environment of the consumer may be appealing to researchers who wish to better understand the consumer choice under consideration.⁵

We show how B2C communication technologies can be used to obtain

³Other lab studies with student samples also tend to report smaller shares of risk-averse subjects than in our field setting with non-students. Gioia (2017) reports that 64.3% of the subjects make decisions in the BRET that are consistent with risk aversion. Basteck and Mantovani (2018) summarize the BRET choices made by subjects in a histogram. The figure reveals that the modal choice was to collect 50 boxes, which is consistent with risk neutrality. They do not report the share of risk-averse subjects.

⁴See Schildberg-Hörisch (2018) for a discussion on individual risk attitudes in traditional models of decision-making.

⁵This also resonates with the call from Barseghyan et al. (2018) for more research that combines experimental and field data to study individual risk preferences and choices.

comprehensive and timely individual risk profiles in field settings. The ability to implement incentive-compatible experimental methods on the fly changes the aforementioned trade-off with survey-based approaches. We show that it is now possible to implement both approaches alongside each other in what were once traditionally offline consumer settings. The richer estimates of individual risk attitudes that follow from this may significantly improve our understanding of consumer choices in these settings.

This chapter proceeds as follows. Section 5.2 discusses related literature. Section 5.3 describes the research design and elaborates on the field setting. Analyses of the data from the field study are presented in Section 5.4. Section 5.5 concludes the chapter.

5.2 Related literature

Individual risk attitudes are at the heart of many decisions. In empirical work on decision-making under uncertainty, researchers therefore wish to have an estimate of these risk attitudes, either to control for it or because of an interest in risk preferences in itself (Crosetto and Filippin 2016). Selecting the appropriate risk elicitation method depends on cost- and time considerations, subjects' numeracy skills, and the required infrastructure. Incentive compatibility and precision of the elicited risk attitudes are other factors taken into consideration. Depending on relative weights assigned to each consideration, researchers then generally opt for either a survey-based approach or an experimental approach to risk preferences elicitation.

Survey questions about risk attitudes typically ask individuals to make a global assessment of their willingness to take risk, as with the SOEP risk question, or to make domain-specific assessments of individual risk attitudes, as is done in the Domain-Specific Risk-Taking Scale (DOSPERT, Blais and Weber 2006). Responses to the SOEP risk question and the overall DOSPERT score are highly and significantly correlated (Crosetto and Filippin 2016). Both measures tend to correlate well with actual risk-taking behavior. Dohmen et al. (2011) find a positive correlation between willingness to take risks in general (as measured by the SOEP question) and risky behaviors, such as investments in stocks, smoking, and being self-

employed. Hanoch, Johnson, and Wilke (2006) observe that the DOSPERT scale correlates well with risky behaviors within specific domains and conclude that risk preferences may be domain-specific. Despite the practical advantages and the apparent good performance of survey-based measures, there continues to be a concern that the non-incentivized nature of the survey questions may distort reported risk attitudes due to factors like self-serving biases and inattention (Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner 2011). Experimental methods aim to address this concern.

Experimental methods elicit risk attitudes in an incentive-compatible way by letting subjects work on choice tasks in which real money is at stake. Well-known experimental elicitation tasks are the Holt-Laury Multiple Price List (Holt and Laury 2002), the Gneezy-Potters Investment Game (Gneezy and Potters 1997), the Eckel-Grossman Ordered Lottery Task (Eckel and Grossman 2002), the Balloon Analogue Risk Task (BART) (Lejuez et al. 2002), and the Bomb Risk Elicitation Task (BRET) (Crosetto and Filippin 2013). Several studies have appeared that evaluate the relative merits and drawbacks of these different experimental methods (Charness, Gneezy, and Imas 2013, Dave, Eckel, Johnson, and Rojas 2010, Cox and Harrison 2008). Some of the tasks are comparatively simple, either because they have a visual design that makes them easy to understand (e.g., BART and BRET) or they require only a single choice from subjects (Investment Game and the Ordered Lottery Task). Other tasks, such as those involving multiple price lists, are more complex and may yield richer but possibly also noisier data on individual risk attitudes.⁶

Due to the cost and complexities involved, existing studies with experimental elicitation tasks tend to be restricted to laboratory settings with student populations or to lab-in-the-field studies in developing countries.⁷ To the best of our knowledge, we are not aware of any study in a developed

⁶As noted by Charness et al. (2013), a downside of more complex elicitation tasks is that possibly a non-negligible share of subjects will fail to comprehend the task. For example, Charness and Viceisza (2016) administer the Holt-Laury task and find that more than half of the subjects make inconsistent choices. Jacobson and Petrie (2009) report that 55% of their subjects make inconsistent choices in sequential choice lotteries.

⁷Two notable exceptions are von Gaudecker, van Soest, and Wengstrom (2011) and Harrison, Lau, and Rutström (2007). These studies use experimental tasks to elicit the risk attitudes of representative populations in the Netherlands and Denmark, respectively.

country that uses experimental tasks to elicit the risk attitudes of a large representative population in a natural field setting (wherein subjects are not aware of partaking in an experiment). Our study aims to achieve this.

A recurrent finding in the literature is that the correlation between choices in experimental tasks and survey-based measures of risk attitudes is low (Crosetto and Filippin 2016). Charness and Viceisza (2016) and He (2018) administer both the SOEP question and incentivized risk elicitation tasks to subjects from rural populations in respectively Senegal and China and report that there is no significant correlation. Dohmen et al. (2011) do find a significant relation between the SOEP question and an incentivized lottery task, but the variance explained is low ($\approx 6\%$). Deck et al. (2013) report that differences in subjects' behavior across several experimental tasks cannot be explained by differences in responses in the relevant domain of the DOSPERT. We find in our field setting a positive but small correlation between choices in the BRET and responses to the SOEP risk question.

In terms of predictive power for actual risky behaviors, risk attitudes elicited through experiments perform relatively poorly in existing studies (Crosetto and Filippin 2016, Charness, Gneezy, and Imas 2013). This may in part be due to the stylized environment in which these experiments are administered, the reliance on student samples, and the fact that these behaviors are typically self-reported measures. The predictive power of estimated risk attitudes may improve when the elicitation task is integrated in the natural environment in which real-life financial decisions by a non-student population are made. In this study, we show how this can be done.

5.3 Research design

This section outlines the research design. We start by discussing the Bomb Risk Elicitation Task (henceforth BRET) in more detail and motivate our design choices by comparing these to those made by Crosetto and Filippin (2013) in their lab tests of the BRET. Next, we discuss the main research objectives and elaborate on the field setting. The section concludes with the implementation plan and a description of the subject pool.

5.3.1 The bomb risk elicitation task (BRET)

The BRET, as developed by Crosetto and Filippin (2013), can be described as follows.⁸ Subjects are presented with 100 boxes and have to decide how many of them to collect. The payment is linearly increasing in the number of boxes a subject collects. There is one caveat: one randomly selected box out of the 100 boxes contains a bomb. If the box that contains the bomb is in the set of collected boxes, the subject will earn nothing. Only at the end of the experiment is it revealed to the subject which box contains the bomb. In the baseline dynamic version of the BRET, boxes are sequentially collected at a speed of one box per second. The subject can at any time decide to stop collecting more boxes by pushing a button. All boxes will then open and reveal whether the bomb is among the ones collected.

An advantage of the BRET is that it offers straightforward predictions about how subjects with different risk attitudes should behave in the task. A risk neutral subject should choose to collect exactly fifty boxes, while risk averse (risk loving) subjects should choose to open less (more), see Crosetto and Filippin (2013) for formal derivations and numerical examples. The number of boxes collected provides a precise estimate of a subject's coefficient of risk aversion r . Other advantages of the BRET, compared to other risk elicitation methods, are that it is exclusively framed in terms of gains (loss aversion cannot come into play), it is fast and easily explained, and the task requires minimal numerical skills. These advantages make the BRET very suitable for field settings.

In their lab studies, Crosetto and Filippin (2012) compare choices in the BRET with responses to a non-incentivized survey question on risk attitudes in the German Socio-Economic Panel Study (SOEP). They find that there is a positive and significant relation between choices in the BRET and responses to the SOEP question. The significance, however, is only at the 10% level and the variance explained is low. It is therefore not entirely

⁸In a between-treatment design, Crosetto and Filippin (2013) run various versions of the task that differ in the presence/absence of a trial period, speed of the deletion process, order in which the boxes are deleted, the size of the boxes etc. For validation, they compare the outcomes with the outcomes of three other elicitation tasks: the Multiple Price List method (Holt and Laury 2002), the Ordered Lottery Choice method (Eckel and Grossman 2002) and the Investment Game (Gneezy and Potters 1997).

clear how these results would generalize to a non-student population outside of the lab. Among other things, we aim to address this concern.

5.3.2 Research objectives

Our main goal in this chapter is to estimate individual risk attitudes of a large non-student population in a natural field setting. We use an incentive-compatible experimental method (BRET) and a non-incentivized survey question (SOEP) to achieve this goal. In addition, we have three subgoals. First, we wish to get an idea of the correlation between choices in the BRET and responses to the SOEP question when both are elicited in the field. Such information is instructive for researchers interested in relating elicited risk attitudes to actual choices in the field.

Second, we consider the stability of the elicited risk attitudes. In our field setting, we have the opportunity to repeatedly administer both elicitation methods in the same population and to track choices and responses through time. This allows us to see at the individual level whether decision patterns in the beginning change or remain stable in subsequent rounds.

Third, many studies find that women are more risk averse than men (Croson and Gneezy 2009). Crosetto and Filippin (2013), however, observe no gender difference in BRET choices. Using customer profile data from our field partner, we examine whether a gender difference is observed in the BRET and in the SOEP question in our non-student subject pool. In a more exploratory analysis, we will also explore the relation between individual risk attitudes and other demographic variables (age and wealth).

5.3.3 Field setting

We collaborate with TanQyou (henceforth TQ), a startup that entered the Dutch market for retail fueling in 2015. The company operates about 20 filling stations, of which 13 are equipped with technologies that enable on-site personal communication with customers via the displays (“Hi Gert-Jan!”). These 13 stations are geographically dispersed in the Netherlands and will constitute the sample of this study.⁹ TQ aims to establish a personal re-

⁹The geographical locations of these filling stations can be found in Table 5.A.1.

lationship with the customer using a wide array of digital technologies. Visitors to TQ's filling stations are invited to register as a member. For every liter of fuel bought during a visit, members receive a ticket for a monthly prize draw. Prizes include travel and dinner vouchers, television screens, smart watches and occasionally even a car. Each station has its own prize pool. Other than on the physical platforms, TQ also interacts with its members through email and an app. Through these communication channels, members have access to other car-related services in TQ's product range, varying from an online diary of fuel expenditures to parking.

TQ is an interesting partner for the purposes of this study because of its digital approach to maintaining customer relationships. This approach makes it possible to uniquely identify members and to administer the BRET in a natural field setting over an extended period of time. The technologies involved ensure that everything is meticulously recorded. We will know of each member when exactly the task was performed and the time spent on each part. The same holds for the SOEP risk question. This information can be matched with member-related data, such as demographics, fuel transactions, registered vehicle, and so on.

In this study, we will restrict attention to the choices and behaviors of TQ members for three practical reasons. First, members can be uniquely identified through their membership IDs. Second, the ability to identify members during repeat visits enables tracking through time. Third, for budgetary reasons, the large sample size requires that the incentives in the BRET are related to services that are exclusively available to members (henceforth referred to as subjects).

5.3.4 Implementation plan and design choices

The pump displays at the filling stations are used to administer the dynamic version of the BRET as well as the SOEP question to visitors who have just completed filling up the tank, conditional on the visitor being a member.¹⁰ Figure 5.A.1 shows an on-site pump display. We use the dynamic version of the BRET because, as shown by Crosetto and Filippin (2013), this version

¹⁰The software makes use of the module designed by Holzmeister and Pfurtscheller (2016) for use with *oTree* (Chen, Schonger, and Wickens 2016).

more likely produces reliable outcomes, especially among subjects with low numeracy skills, than a pen-and-paper static version in which subjects are simply asked how many boxes they wish to collect.

Both the BRET and the SOEP question will be posed once per month to subjects for the duration of one year.¹¹ This allows us to see to what extent choices change or remain stable over time. Absent any wealth effects, theory predicts that subjects should exhibit the same behavior in a repeated version of the BRET as in a one-shot BRET. When the task is repeated, Crosetto and Filippin (2013) find that average behavior stays the same when choices are aggregated, but that there is sizeable within-subject variation in choices. Our main analyses will focus on the first time that subjects play the BRET and answer the SOEP question, but we also consider choices in repeated rounds.

The procedure at the filling station is as follows. Subjects first fill the tank. As discussed in Section 5.3.3, every liter of fuel bought yields a ticket for a monthly prize draw. While filling the tank, the pump display shows in real time how many tickets are earned. Once the filling process is completed, the subject turns to the pump display to finish the transaction. At that moment, the display invites the subject to a “game” in which something extra can be earned. The game being the BRET.

There are two types of earnings, each with its own version of the BRET. In the main version, subjects can earn extra tickets (one per box) for the prize draw. A benefit of this version is that many subjects can be reached without incurring excessive implementation costs; the marginal cost of handing out extra tickets is practically zero. A possible concern, however, is that the main version generates compound lotteries, which may potentially influence choices and/or the willingness to play the game. In a robustness check, discussed in Section 5.4.2, we therefore also once administer an alternative version of the BRET in which the subjects can earn a discount (€0.05 per box) that will be subtracted from the transaction

¹¹We did not administer the DOSPERT because in our field setting that would require too much time from subjects and would disturb the traffic flow at the filling stations. We do believe that it would be interesting in future research to embed the elicitation of DOSPERT responses in natural field settings, especially when behaviors and outcomes in these settings can be correlated with the corresponding domains of the DOSPERT.

amount. This way, we avoid the creation of compound lotteries because the discount is guaranteed as long as the bomb is not among the collected boxes. A quasi-random sample of 600 subjects will perform this version of the BRET. A benefit of both versions is that they are naturally embedded in the field setting. That is, subjects are already familiar with the tickets for prize draws and discounts at the pump.

In both versions, subjects are presented 100 boxes (10x10) that can be collected in sequential order at a speed of 0.4 seconds per box. There is no trial round.¹² Compared to the baseline dynamic version in Crosetto and Filippin (2013), we have sped up the BRET (from 1.0 to 0.4 seconds) and excluded the trial round. This is because we want to minimize the hassle for subjects and be of no disturbance to the traffic flow at the filling stations. In robustness checks, Crosetto and Filippin (2013) find that these two adjustments can safely be made, as stopping behavior is not significantly affected when the task speed increases or when the trial round is absent.¹³

The default language in which task instructions are shown is Dutch. Subjects have the choice to change the language to English or German. The instructions are written to fit the screen of the pump display. This required from us to write the instructions as short as possible while keeping the core message intact. The exact wording used in both versions is as follows (the English translation is shown for convenience)¹⁴:

¹²With respect to offering subjects trial periods, Crosetto and Filippin (2013) worry that subjects' experiences in a trial period may influence their behavior in the for-pay stage. Providing feedback on the position of the bomb may for example create focal points or arouse mean-reversion strategies. Not providing feedback on the position of the bomb may generate feelings of 'having beaten the bomb' (Crosetto and Filippin 2013, p. 48) leading to a higher taste for risk in the for-pay stage. They show that the presence or absence of a trial stage does not result in statistically significant differences in decisions. Given this similarity in outcomes and in light of the potential problems trial stages introduce, we decided not to add a trial round.

¹³Crosetto and Filippin (2013) consider the case where a box is collected every half a second. With 0.4 seconds, we stay close to this speed level.

¹⁴The Dutch and German translations can be found in Appendix section 5.A.3.

Instructions BRET

The game has 100 boxes. 99 boxes contain a ticket but one random box contains a bomb.

When the game starts you collect a box every 0.4 seconds until you press the stop button.

Earn tickets by collecting boxes, but if you collect the bomb, you will lose all the tickets you have collected with this game (you keep the tickets you have earned with refueling).

The SOEP general risk question

One last question: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

0 = Not at all willing to take risks

10 = Very willing to take risks

Figure 5.1 shows the different stages of how the BRET and the SOEP risk question are visualized on the pump display.¹⁵ Right after refueling the tank, subjects are asked via a pop-up screen whether they would be interested in playing a game to earn extra tickets in less than a minute. We include the ex ante time indication to rule out that subjects finish the game early because of time considerations, which we expected might otherwise be more of a concern in the field than in lab settings. A subject is shown the pop-up screen once a month, conditional on not having three times indicated in the past that he or she does not want to play the game. A benefit of the current design is that subjects need to push a button to proceed to the next stage. Since a timestamp is created at every push, we know exactly how much time is spent on each part.

5.3.5 Timeline of the study

Figure 5.2 depicts the timeline of the study. On 26 February 2019, we started with a four-month pilot study at one filling station of our field partner to test the infrastructure and to detect possible errors in data

¹⁵Figure 5.A.2 shows the on-site visualization of the BRET on the pump displays.

Figure 5.1: The BRET and the SOEP Risk Question on the Pump Display

Notes: Visualization of the Bomb Risk Elicitation Task (BRET) and the general risk question from the Socio-Economic Panel Study (SOEP) on the on-site pump displays at our field partner's filling stations in the Netherlands. Instructions shown are in Dutch.

Figure 5.2: Timeline of the Study

recording and the display of the elicitation tasks. No such errors were found. The only adjustment made was the introduction of the pop-up screen (panel A in Figure 5.1) on 24 April 2019. This was because the original button was not clearly visible on the screen and therefore often undetected by subjects. Since no adjustments were made in the instructions, nor in the visual layout of the elicitation methods, we decided to keep the data from the pilot phase in the sample used for the empirical analyses. On 27 June 2019 the BRET and the SOEP question were rolled out over the 11 remaining filling stations with a digital infrastructure. In late December 2019, a new filling station was included in the sample, bringing the total number of participating filling stations to 13.

Due to the COVID-19 pandemic, the robustness check in which subjects complete the “discount” version of the BRET is, at the time of writing, postponed until the fall of 2020. This is because of reduced activity at the filling stations following the lockdown measures in the Netherlands.

5.3.6 Data description

The full sample period for the baseline BRET version is from 26 February 2019 until 30 June 2020. Data are available until 19 February 2020, the last date of data sharing with our field partner prior to the COVID-19 outbreak. Up till this date, 3,870 members of TQ received an invitation via the on-site pump display to complete the BRET and the SOEP general risk question. In total, 2,439 members (63%) completed the BRET at least once, of which 2,143 (88%) additionally also answered the SOEP general risk question at least once. As shown in Figure 5.A.3, the average weekly participation rate at filling stations fluctuates around 25% and is quite stable across time. From the point of view of survey design in natural field settings, the stable

participation share over an extended period of time is an encouraging sign that a large share of the intended sample can be reached through innovative on-site communication technologies.

In our presentation of the results and descriptive statistics, we exclude subjects who made obvious mistakes in the BRET. We follow Crosetto and Filippin (2013) and label choices as obvious mistakes when subjects stopped at $k = 0$, $k = 1$, or $k = 100$, where k denotes the number of boxes collected in the BRET. Stopping very early in the task is likely by mistake and therefore does not result in a reliable estimate of risk attitudes. Collecting 100 boxes is a dominated choice because it results in a loss of earnings with probability one, and hence suggests a failure of understanding the task. In the first round of the BRET, excluding obvious mistakes led to the removal of 286 observations (11.73%), and in subsequent rounds it led to the exclusion of another 145 observations (4.78%). Hence, of all submitted choices in the BRET (5,743 in total), 431 observations (7.5%) were labeled as obvious mistakes and removed from the dataset.

Table 5.1 summarizes the demographics of the subjects in the experimental sample and contrasts these with the full population demographics of the municipalities where the filling stations are located. About 69% of the subjects are male. All age groups between 18 and 80 are well-represented in the sample, providing a rich and wide cross-section of the underlying population. The average age of the subjects is 46 years old.¹⁶ Qualitatively, when compared to the age distribution of the full population in the sample area, the relative sizes of the age groups in the sample are very similar. Being able to reach such a diverse group of subjects is a key advantage of eliciting risk attitudes in a natural field setting like filling stations.

¹⁶In the Netherlands, the legal driving age is 18.

Table 5.1: Demographics of the Experimental Sample

	Age bracket						<i>N</i>
	15-24	25-44	45-64	65-79	80 plus	N.A.	
Male	39	254	357	83	1	762	1,496
Female	15	145	147	10	0	340	657
Total	54	399	504	93	1	1,102	2,153
% share of total	5.14	37.96	47.95	8.85	0.1		
% share female	27.78	36.34	29.17	10.75	0.00		
Full population in the municipalities of the filling stations							
Total	79,954	151,974	195,496	105,016	32,971		
% share of total	13.85	26.08	35.08	19.04	5.95		

Notes: Data on the full population of the municipalities where the filling stations are located are retrieved from Statistics Netherlands (CBS). The legal driving age in the Netherlands is 18 and this is therefore the minimum age of subjects in the experimental sample. The age category 15-24 is used, however, to facilitate the comparison with the available data from CBS. The reported percentage shares of the full population are the averages of the shares in the municipalities.

5.4 Results

5.4.1 Main results

The distribution of boxes collected in the first round of the BRET (i.e., the first time subjects completed the task) is shown in Figure 5.3. The figure clearly shows a tendency of subjects to collect boxes in a way that is consistent with risk aversion. To be precise, about 82% of the subjects classify as risk averse ($k \leq 49$), 1% as risk neutral ($k = 50$), and 17% as risk loving ($k \geq 51$). Splitting the data by quartiles, we observe that 50% of the observations lie between 15 and 43. First-round decisions in the task are also remarkably stable through time, as indicated in Figure 5.A.4, with weekly averages fluctuating in a narrow range around 30 boxes collected.¹⁷ Compared to the BRET results from the lab reported in Crosetto and Filippin (2013)¹⁸, we find that in the field a significantly larger share of subjects is risk averse (an increase of about 30 percentage points).

¹⁷The decisions are also stable across the filling stations included in the sample. In an additional analysis (not shown) we regress first-round decisions in the BRET on location fixed effects and find no significant differences between the stations. The exception is the filling station in Westergeest, where the number of boxes collected is significantly lower, but we stress that this result is based on a very small sample of only three subjects.

¹⁸They find that 52.1% of the subjects are risk averse, and that 14.7% (33.2%) are risk neutral (risk loving).

Figure 5.3: Distribution of Boxes Collected in the BRET

Notes: The figure is based on first-round decisions in the Bomb Risk Elicitation Task (BRET).

Figure 5.4: Distribution of the Risk Aversion Parameter r

Notes: Based on the CRRA utility function $u(k) = k^r$, where k is the number of boxes collected in the Bomb Risk Elicitation Task (BRET). For ease of illustration, the range of parameter values shown is restricted to $r \leq 3$. The red dotted line indicates the risk aversion parameter of risk neutrality ($r = 1$). Parameter values smaller (greater) than $r = 1$ indicate risk aversion (risk loving). The figure is based on first-round choices in the BRET.

Assuming a CRRA utility function¹⁹, the number of boxes collected can be used to compute for each subject the risk aversion parameter r . With $r = 1$ as the baseline for risk neutrality, smaller (larger) values of r indicate behavior consistent with risk aversion (risk seeking). Figure 5.4 shows the distribution of r in our field setting and reveals that the majority of subjects have parameter values in line with risk aversion.

The first-round responses to the SOEP risk question are shown in Figure 5.5. The modal response value of seven suggests that subjects are moderately willing to take risks. That the value zero is never chosen indicates that all subjects are at least willing to take some risk. At first sight, the distribution of responses to the SOEP risk question paints a different picture than the outcomes observed in the BRET. Further analyses, however, show that the correlation between the BRET and the SOEP is positive and significant ($p = 0.001$), albeit small in magnitude.²⁰ The variance explained is also low ($R^2 = 0.006$). The observed low correlation between the outcomes of the two risk elicitation methods is in line with previous studies, as discussed in Section 5.2. There is thus some, albeit weak, evidence that subjects in our field setting who choose to open fewer (more) boxes in the BRET also tend to select a lower (higher) value in the SOEP risk question.

We find no gender differences in the decision how many boxes to collect in the first round of the BRET. Table 5.2 shows that the average of boxes collected is similar for males and females (31 and 32 respectively). Interestingly, when subjects are asked to rate their willingness to take risk, as in the SOEP question, we do observe a significant gender difference, with males being more willing to take risks than females.²¹ Dohmen et al. (2011) and

¹⁹ $u(k) = k^r$, where k is the number of boxes collected in the BRET and r is the risk aversion parameter.

²⁰A one-unit increase in the SOEP score is associated with collecting 0.78 extra boxes in the BRET. Spearman's rank correlation coefficient between BRET choices and SOEP responses in the first round is $\rho = 0.0657$ ($p = 0.005$). Figure 5.A.5 shows a scatter plot of the relation between the elicited risk aversion parameter in the BRET and responses to the SOEP risk question.

²¹Since responses to the SOEP general risk question are measured on an ordinal scale, Crosetto and Filippin (2013) do not compare mean responses to the SOEP question but instead transform the ordinal scale into a dummy variable. They chose a cut-off point, the value six, such that the percentage of risk-averse subjects and risk-loving subjects is similar to the shares found in the BRET. The gender difference in responses to the

Figure 5.5: Responses to the SOEP Risk Question

Notes: First-round responses to the SOEP question on willingness to take risks in general. Responses are measured on an eleven-point scale from 0 (not at all willing to take risks) to 10 (very willing to take risks).

Table 5.2: First-Round Gender Differences in Elicited Risk Attitudes

	BRET			SOEP		
	<i>N</i>	Mean	<i>p</i> -value	<i>N</i>	Mean	<i>p</i> -value
Male	1,496	30.96	0.661	1,233	6.52	0.001
Female	657	31.89		556	6.20	
Total	2,153	31.24		1,789	6.42	

Notes: Reported *p*-values are based on Wilcoxon-Mann-Whitney tests.

Crosetto and Filippin (2013) also find that female subjects are more risk averse in the SOEP risk question. The latter study puts forward the idea that the observed discrepancy may be driven by loss aversion and not so much by differences in risk aversion. In a robustness check, they present evidence in support of this claim.²² Given the scope of this chapter, we did not run a separate loss-aversion treatment, but we do encourage other researchers to investigate in natural field settings what exactly gives rise to the frequently observed gender differences in risk preferences.

Other demographic factors do not explain any of the choice heterogeneity in the BRET. In Table 5.A.2 we regress choices in the BRET on gender, age, and the current vehicle value (used as a proxy for wealth). We find that none of these demographic factors significantly influence the choice how many boxes to collect in the BRET. Dohmen et al. (2011) report that the willingness to take risk decreases with age, whereas Harrison et al. (2007) find in a representative sample of the Danish population that in particular middle-aged individuals (those aged between 40 and 50) are less risk averse than younger and older age groups. No such age effects are observed in our sample. Dohmen et al. (2011) also indicate that wealthier individuals are more willing to take risks, though they exercise caution in assigning a causal interpretation to this result due to potential endogeneity issues between risk attitudes and wealth accumulation. A positive relation between wealth and willingness to take risk is not observed in our data. The estimates in Table 5.A.2 do suggest that time spent on reading the BRET instructions is positively associated with the number of boxes collected in the task.

SOEP question, as mentioned in the main text, is still observed in our data if we use this alternative procedure. Results are not shown for brevity but are available upon request.

²²The idea is that female subjects possibly attach more weight to losses and may therefore be drawn to safer choices in which such losses are less likely to occur. In the standard version of the BRET, which is entirely defined in the gain domain, loss aversion is ruled out by design and this may explain why no gender difference is observed. In the robustness check, Crosetto and Filippin (2013) use a modified version of the BRET in which a reference point is explicitly induced. They find suggestive evidence that gender differences in loss aversion, triggered by the reference point, lead to gender-specific differences in the number of boxes collected in the BRET.

5.4.2 Additional analyses and robustness checks

Results for the BRET as a simple lottery The previous results are based on a version of the BRET in which tickets for prize draws are used as the incentive. As discussed in Section 5.3.4, this incentive was chosen to avoid excessive implementation costs, but also because subjects in our field setting are already familiar with the tickets. This way, we were able to work with succinct instructions that limit the disturbance to the traffic flows at the filling stations and subjects' time schedules. A possible concern, however, is that the lottery tickets create compound lotteries in the BRET and that this, in turn, may influence the willingness to participate in the task and/or choices in the task itself. Kaivanto and Kroll (2012), for example, find that subjects are more risk averse when the gamble is framed as a compound lottery rather than as a simple lottery.

To see to what extent our results would change if the BRET is instead presented as a simple lottery, we let a quasi-random sample of 600 subjects perform a version of the BRET in which they could earn up to €5 as a discount that will be deducted from the fuel transaction amount shown on the pump display. Each box in the BRET thus contains €0.05. The discount is guaranteed as long as the bomb is not among the collected boxes in the task, thereby avoiding the issue of compound lotteries.

The implementation of this robustness check is postponed because of the COVID-19 outbreak in the Netherlands. The simple lottery version of the BRET was initially scheduled to be implemented in April and May 2020. At the time of writing, this is postponed until the fall of 2020. We are currently in close consultation with our field partner to determine what is safe and technically feasible at the filling stations in order to be able to select a quasi-random sample of 600 subjects around that new period.

Attention during the task One concern of eliciting risk attitudes in the field is that, compared to a lab setting, subjects may have less time and attention available for the elicitation tasks. Comprehension difficulties may follow when subjects do not carefully read the instructions, and this, in turn, may impact our elicited risk attitudes. To assess the magnitude of this issue, we added timestamps when a subject pushed a button to

continue to the next screen and to proceed with the next stage of the tasks (see Figure 5.1). By determining the elapsed time between the pushing of buttons, we can measure precisely how much time is spent on each part.

Figure 5.A.6 presents the results of this exercise.²³ The figure shows per round how many seconds, on average, were spent on (1) reading the instructions of the BRET, (2) performing the BRET plus observing the outcome, and (3) answering the SOEP general risk question. In the first round, subjects needed about eight seconds to read the instructions of the BRET and another 18 seconds to perform the BRET and observe the outcome. An additional ten seconds were needed to submit a response to the SOEP question. In subsequent rounds, the time spent on the BRET instructions and the SOEP question appears to be steadily declining, as can be expected, whereas the time used for performing the BRET remains stable at around 18-19 seconds. Since these numbers appear very reasonable, given the content size at each stage, we conclude that subjects paid careful attention to the tasks displayed on the pump screen.

Wealth effects In the baseline version of the BRET we use tickets for periodic prize draws that are organized by our field partner. Each box collected in the task corresponds to one extra ticket, provided the bomb is not in one of the collected boxes. These tickets are earned on top of the tickets earned in the process of filling the tank. As mentioned in Section 5.3.4, the company's policy is to give filling station visitors one ticket per liter of fuel bought, and this is made salient to the visitors via a real-time counter on the pump display. One potential concern is that wealth effects in this setting may influence choices in the BRET. In their validation tests of the BRET, Crosetto and Filippin (2013) test for wealth effects by letting subjects first complete another task, the BART, and then track whether earnings in this task affect subsequent choices in an unannounced version of the BRET. They find that prior earnings lead to convex wealth effects in the BRET.

In our setting, the first round of the BRET can be treated as unannounced as well, since we did not inform the subjects about the phase-in

²³We only show the results for the first ten rounds due to the very limited number of subjects (5 in total) who completed more than ten rounds.

of the BRET at the filling stations. We can therefore rule out that in the first round subjects altered their filling behavior, and thus the number of “filling tickets” collected, in anticipation of performing the BRET afterwards. Regressing the number of boxes collected in the BRET on the number of filling tickets obtained just prior to the task therefore gives an indication of possible wealth effects. The regression results are reported in Table 5.A.3 and show no sign of linear or non-linear wealth effects. The presented estimates of risk attitudes therefore do not seem to be affected by prior “earnings” from other activities in our field setting.

Repeated rounds of the BRET and SOEP The on-site technology at the filling stations makes it possible to repeatedly administer the same risk elicitation tasks to the same group of subjects over a long time period. As discussed in Section 5.3.4, each subject is invited about once a month via the pump display to partake in the tasks. By collecting the responses in each round, we can examine whether risk attitudes remain stable or change in the two tasks under consideration. Research on the stability of risk preferences is important for policy design and for improving our understanding of intertemporal decision-making (Schildberg-Hörisch 2018).

Table 5.A.4 indicates that, compared to the first round, the average number of boxes collected in the BRET is significantly higher in subsequent rounds. The table focuses on choices made by subjects who participated in all of the first five rounds in order to keep the composition of the subject pool in each round constant.²⁴ The increase in the number of boxes collected in the task suggests that subjects become more willing to take risks when the task is performed multiple times. This may reflect time-varying individual risk attitudes, although we do exercise caution in interpreting this result because of the possible serial correlation between rounds.²⁵ For

²⁴Table 5.A.6 provides the descriptive statistics of BRET choices in all rounds.

²⁵Another explanation is that subjects have a fixed amount of time available for reading the BRET instructions and completing the task. In repeated rounds, subjects may need less time for the instructions and thus have more time available for the task itself. In this case, we would expect to see that the decrease in time spent on the instructions is exactly offset by the increase in time spent on the BRET. We find no evidence of this. The difference between the change in time spent on the instructions and the change in time spent on the task itself is significantly different from zero ($p < 0.001$).

example, subjects may be prone to a form of gambler fallacy and take the position of the bomb in the previous round into consideration when deciding how many boxes to collect in the current round. We find some evidence in line with this. The correlation between boxes collected in a round and the bomb position in the previous round is 0.220 ($p < 0.001$).

To investigate the stability of individual BRET choices further, we follow the definition of Crosetto and Filippin (2013) and classify a subject as having stable risk preferences when for that subject the maximum and minimum number of collected boxes in repeated rounds does not exceed ten ($\max(k) - \min(k) \leq 10$). According to this definition, and by restricting attention again to the group of subjects who participated in all of the first five rounds, 2.27% of the subjects have stable risk attitudes over time. This share is notably smaller than the 29.51% reported by Crosetto and Filippin (2013). In their design, however, the position of the bomb is only revealed at the end of all rounds, when one randomly determined round is selected to be payoff-relevant. This procedure is less feasible for field settings, including ours, because of the temporal distance between rounds.

Responses to the SOEP risk question in repeated rounds also show that subjects display a tendency to become more willing to take risks over time. Since the risk question is measured on an ordinal scale, we consider the risk attitudes of a subject to be stable in a given round when the response to the risk question is equal to the response in the first round (e.g., submitting the value seven in both rounds). Table 5.A.5 considers the responses of subjects who responded to the SOEP question in all of the first five rounds.²⁶ We observe that 34.55% of the responses can be classified as “stable”. In contrast, 35.96% (29.49%) of the responses are in line with an increased (decreased) willingness to take risks in repeated rounds. Compared to the first round, 15.17% (9.55%) of the subjects consistently submit a larger (smaller) value. The share of subjects who always choose the same answer is 9.55%. The remaining subjects (65.73%) pick answers that are sometimes larger and sometimes smaller than the response submitted in the first round.

²⁶A complete overview of the responses to the SOEP risk question in all rounds can be found in Table 5.A.7.

5.5 Conclusion

Individual risk attitudes are at the heart of many decisions that involve uncertainty, and its measurement has therefore attracted much attention. A challenge identified in the literature is how to elicit risk attitudes in an incentive-compatible way in the field, where computer access is limited. Particularly in settings where risk attitudes are domain-specific and time-varying, having access to timely and incentive-compatible estimates of risk attitudes may prove invaluable for studying individual choices and actions.

In this chapter, we argue and show that innovations in business-to-consumer (B2C) communication increase the scope for the incentivized measurement of individual risk attitudes in the field. We partner with a startup in the retail fuel industry and for one year use their novel on-site pump displays at 13 geographically-dispersed filling stations to repeatedly administer two risk elicitation methods to a large and diverse population of about 2,500 gas station visitors: the incentive-compatible Bomb Risk Elicitation Task (BRET) and the non-incentivized survey question on willingness to take risk from the German Socio-Economic Panel Study (SOEP).

We find that 82% of the subjects in this natural field setting make decisions in the BRET that are consistent with risk aversion. This is about 30 percentage points higher than the share reported by Crosetto and Filippin (2013) in their laboratory tests of the BRET with a student sample. Second, no gender differences are observed in BRET choices, nor do other demographics explain some of the variance in the outcomes of the task. Third, the correlation between BRET choices and responses to the SOEP question is positive and significant, but small in magnitude. Fourth, in repeat elicitations, we find that risk attitudes tend to vary over time. Finally, our approach shows that, even with voluntary participation, a large share of the intended sample in the field can be reached through communication technologies that display the elicitation methods during a wide time frame.

The advent of digital communication technologies in consumer markets makes it possible to elicit risk attitudes on the fly in field settings with large and diverse populations, thereby offering new ways to investigate how risk attitudes shape decisions. In future work, we aim to explore this further.

5.A Appendix

5.A.1 Geographical locations of the filling stations

Table 5.A.1: Filling Stations in the Sample

Location	Province	Start Data Recording
Appelscha	Friesland	27 June, 2019
Bergen op Zoom	Noord-Brabant	27 June, 2019
Bleiswijk	Zuid-Holland	27 June, 2019
Dalfsen	Overijssel	27 June, 2019
Franeker	Friesland	26 February, 2019
Grijpskerk	Groningen	27 June, 2019
Heerlen	Limburg	27 June, 2019
Landgraaf	Limburg	27 June, 2019
Leeuwarden	Friesland	27 June, 2019
Ommen	Overijssel	27 June, 2019
Putten	Gelderland	27 June, 2019
Sint Annaparochie	Friesland	27 June, 2019
Westergeest	Friesland	20 December, 2019

5.A.2 Pump displays

Figure 5.A.1: Pump Display at Filling Station

Figure 5.A.2: On-Site Visualization of the BRET on Pump Displays

5.A.3 Instructions in Dutch and German

Instructions BRET - Dutch

-Playing for tickets-

Het spel telt 100 dozen. 99 dozen bevatten een lot, maar één willekeurige doos bevat een bom.

Wanneer het spel start, verzamel je iedere 0.4 seconde een doos totdat je de stop-knop indrukt.

Verdien loten door dozen te verzamelen, maar pak je de bom, dan verlies je alle loten die je met dit spel hebt verzameld (de loten die je hebt verdiend met tanken behoud je).

The SOEP general risk question - Dutch

Nog een laatste vraag: zie je jezelf in het algemeen als iemand die bereid is om risico's te nemen of als iemand die risico's probeert te vermijden?

0 = Totaal risicomijdend

10 = Zeer risiconemend

Instructions BRET - German

-Playing for tickets-

Das Spiel hat 100 Boxen. 99 Boxen enthalten ein Lotterielos aber eine zufällige Box enthält eine Bombe.

Wenn das Spiel beginnt, sammeln Sie alle 0,4 Sekunden eine Box, bis Sie die Stopp-Taste drücken.

Versuchen Sie, so viele Lotterielose wie möglich zu sammeln, aber wenn Sie die Bombe bekommen, verlieren Sie alle Lotterielose, die Sie mit diesem Spiel gesammelt haben (Sie behalten die Lotterielose, die Sie beim Tanken gesammelt haben).

The SOEP general risk question - German

Eine letzte Frage: Sind Sie im Allgemeinen ein risikobereiter Mensch oder versuchen Sie, Risiken zu vermeiden?

0 = Gar nicht risikobereit

10 = Sehr risikobereit

5.A.4 Participation share BRET at the filling stations

Figure 5.A.3: Weekly Participation Share BRET at Filling Stations

Notes: The figure shows per filling station the weekly share of subjects who completed the BRET after receiving an invitation via the pump display to participate (pilot phase not included). The participation share for the filling station in Westergest is not shown because the station is new and only a few subjects were eligible for participation.

5.A.5 Choices in the BRET through time

Figure 5.A.4: Boxes Collected in the First Round of the BRET Through Time

Notes: The figure shows the average number of boxes collected in the Bomb Risk Elicitation Task (BRET) per week in the sample period. The sample period started on 26 February 2019 with a four-month pilot phase at one filling station. On 27 June 2019 the BRET was rolled out over the remaining 11 filling stations in the Netherlands. In late December 2019, a new filling station was added to the sample, bringing the total number of participating locations to 13.

5.A.6 Scatter plot risk aversion parameter and responses to the SOEP risk question

Figure 5.A.5: Scatter Plot Risk Aversion Parameter and SOEP Risk Score

Notes: Scatter plot and a fitted regression line of the relation between the risk aversion parameter r (restricted to ≤ 3) and responses to the SOEP risk question. Data are based on first-round choices. The risk aversion parameter r is based on the CRRA utility function $u(k) = k^r$, where k is the number of boxes collected by a subject in the first round of the Bomb Risk Elicitation Task (BRET).

5.A.7 Demographics and choices in the BRET

Table 5.A.2: Boxes Collected in the BRET and Demographics

	(1)	(2)	(3)
Gender	-1.115 (1.958)	-1.146 (1.950)	0.292 (1.449)
Age	-0.268 (0.415)	-0.240 (0.417)	0.051 (0.301)
Age ²	0.002 (0.004)	0.002 (0.004)	-0.001 (0.003)
Current vehicle value	0.104 (0.103)	0.117 (0.102)	0.059 (0.059)
Time spent on instructions		0.217** (0.109)	0.212** (0.085)
Constant	44.849*** (11.461)	42.468*** (11.630)	31.200*** (7.586)
R^2	0.026	0.031	0.018
Number of obs.	619	619	1,844
Round(s)	First	First	All
Round fixed effects	n.a.	n.a.	Yes
Month fixed effects	Yes	Yes	Yes

Notes: Reported estimates are from OLS regressions with the number of boxes collected in the Bomb Risk Elicitation Task (BRET) as the dependent variable. Gender is a dummy variable equal to one (zero) for female (male) subjects. Age is measured in years. The current vehicle value of the subject is calculated as follows (assuming a 15% depreciation rate per annum): $0.85^{VA} \times CP$, where VA denotes vehicle age in years and CP captures the catalogue price of the vehicle in thousands. Time spent on instructions is the time spent in seconds on reading the instructions (see also Section 5.4.2). Standard errors in column (3) are clustered at the subject level.

***(**, *): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

5.A.8 Time spent on tasks per round

Figure 5.A.6: Time Spent on Tasks per Round

5.A.9 Wealth effects in BRET choices

Table 5.A.3: Wealth Effects in the First Round of the BRET

	(1)	(2)	(3)
Filling tickets	0.006 (0.029)	0.016 (0.041)	-0.117 (0.143)
Filling tickets ²		-0.000 (0.000)	0.002 (0.002)
Filling tickets ³			-0.000 (0.000)
Constant	31.038*** (1.090)	30.832*** (1.314)	32.676*** (2.278)
R^2	0.000	0.000	0.000
Number of obs.	2,153	2,153	2,153

Notes: The dependent variable is the number of boxes collected by subjects in the first round of the Bomb Risk Elicitation Task (BRET). Filling tickets are tickets earned while refueling the tank. Each liter of fuel corresponds to one ticket in a prize draw organized by our field partner.

***(**,*): statistically different from zero at the 1%-level (5%-level, 10%-level). Standard errors in parentheses.

5.A.10 Repeated rounds in the BRET and the SOEP

Table 5.A.4: Choices in the BRET in Repeated Rounds

Round	N	Average	S.D.	WMW test p -value
First	220	32.86	19.20	
Second	220	33.99	19.94	0.604
Third	220	35.80	19.35	0.087
Fourth	220	36.28	18.22	0.028
Fifth	220	37.91	19.14	0.004
Second-fifth	880	36.00	19.19	0.021

Notes: Subjects are included conditional on participating in all of the first five rounds. The first round is the first time that subjects performed the Bomb Risk Elicitation Task (BRET). Reported p -values are from Wilcoxon-Mann-Whitney tests that are based on pairwise comparisons of each repeated round with choices in the first round.

Table 5.A.5: Responses to the SOEP Risk Question in Repeated Rounds

Round	N	Stable	Increase	Decrease	χ^2 -test p -value
Second	178	39.89%	29.78%	30.34%	0.075
Third	178	34.27%	36.52%	29.21%	0.326
Fourth	178	30.90%	39.33%	29.78%	0.113
Fifth	178	33.15%	38.20%	28.65%	0.161
Second-fifth	712	34.55%	35.96%	29.49%	0.000

Notes: Subjects are included conditional on responding to the question in all of the first five rounds. The table shows per repeated round the share of subjects who submitted a response that was equal (“stable”) to - or higher/lower than - the response submitted in the first round. The first-round modal response was seven.

Table 5.A.6: Descriptive Statistics of BRET Choices in Repeated Rounds

Round	N	Average	S.D.
First	2,153	31.24	21.08
Second	1,230	32.87	21.13
Third	744	35.34	20.37
Fourth	421	35.62	18.90
Fifth	244	38.05	19.64
Sixth	129	37.44	18.02
Seventh	56	35.84	16.70
Eighth	28	30.89	14.44
Ninth	15	35.07	17.12
Tenth	8	31.00	15.03
First-tenth	5,028	33.16	20.67

Notes: The first round is the first time that subjects performed the Bomb Risk Elicitation Task (BRET).

Table 5.A.7: Responses to the SOEP Risk Question in All Rounds

Round	N	% share of Responses										
		0	1	2	3	4	5	6	7	8	9	10
First	1,931	0.00	1.81	3.37	4.56	5.80	14.91	14.66	24.96	17.71	4.71	7.51
Second	1,053	0.00	1.80	2.47	3.23	5.51	15.19	14.06	26.21	20.80	3.80	6.93
Third	625	0.00	0.80	1.44	4.96	4.48	14.72	14.40	29.12	18.24	4.80	7.04
Fourth	346	0.00	0.58	1.73	3.18	5.20	14.16	13.01	25.43	19.65	5.49	11.56
Fifth	192	0.00	0.00	3.13	3.13	4.17	16.15	11.46	25.00	22.40	4.69	9.90
Sixth	88	0.00	0.00	2.27	4.55	4.55	13.64	13.64	26.14	20.45	5.68	9.09
Seventh	41	0.00	2.44	2.44	0.00	4.88	19.51	12.20	24.39	19.51	9.76	4.88
Eighth	19	0.00	0.00	10.53	10.53	5.26	10.53	5.26	21.05	26.32	5.26	5.26
Ninth	9	0.00	0.00	0.00	11.11	0.00	33.33	0.00	22.22	33.33	0.00	0.00
Tenth	6	0.00	0.00	0.00	0.00	16.67	16.67	0.00	16.67	33.33	0.00	16.67
First-tenth	4,310	0.00	1.44	2.71	4.11	5.38	14.99	14.06	25.89	19.07	4.62	7.73

Bibliography

- ALLCOTT, H. (2011): “Social norms and energy conservation,” *Journal of Public Economics*, 95(9-10), 1082–1095.
- ALLCOTT, H., AND J. B. KESSLER (2019): “The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons,” *American Economic Journal: Applied Economics*, 11(1), 236–276.
- ALLCOTT, H., AND S. MULLAINATHAN (2010): “Behavior and Energy Policy,” *Science*, 327(5970), 1204–1205.
- ALLCOTT, H., AND T. ROGERS (2014): “The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation,” *American Economic Review*, 104(10), 3003–3037.
- ANDERSEN, S., G. W. HARRISON, M. I. LAU, AND E. E. RUTSTRÖM (2014): “Discounting Behavior: A Reconsideration,” *European Economic Review*, 71, 15–33.
- ANDREONI, J., AND C. SPRENGER (2012): “Estimating Time Preferences from Convex Budgets,” *American Economic Review*, 102(7), 3333–3356.
- ANGRIST, J., D. LANG, AND P. OREOPOULOS (2009): “Incentives and Services for College Achievement: Evidence from a Randomized Trial,” *American Economic Journal: Applied Economics*, 1(1), 136–163.
- ARIELY, D., AND K. WERTENBROCH (2002): “Procrastination, Deadlines, and Performance: Self-Control by Precommitment,” *Psychological Science*, 13(3), 219–224.

- ARROW, K. J. (1994): "Methodological Individualism and Social Knowledge," *American Economic Review*, 84(2), 1–9.
- ASHRAF, N., O. BANDIERA, AND S. S. LEE (2014): "Awards unbundled: Evidence from a natural field experiment," *Journal of Economic Behavior and Organization*, 100, 44–63.
- ASHRAF, N., C. F. CAMERER, AND G. LOEWENSTEIN (2005): "Adam Smith, Behavioral Economist," *Journal of Economic Perspectives*, 19(3), 131–145.
- AUGENBLICK, N., M. NIEDERLE, AND C. SPRENGER (2015): "Working over Time: Dynamic Inconsistency in Real Effort Tasks," *The Quarterly Journal of Economics*, 130(3), 1067–1115.
- AYRES, I., S. RASEMAND, AND A. SHIH (2013): "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage," *Journal of Law, Economics, and Organization*, 29(5), 992–1022.
- AZMAT, G., AND N. IRIBERRI (2010): "The importance of relative performance feedback information: Evidence from a natural experiment using high school students," *Journal of Public Economics*, 94(7-8), 435–452.
- AZOULAY, P., J. S. G. ZIVIN, AND J. WANG (2010): "Superstar Extinction," *Quarterly Journal of Economics*, 125(2), 549–589.
- BABCOCK, P., AND M. MARKS (2011): "The Falling Time Cost of College: Evidence from Half a Century of Time Use Data," *The Review of Economics and Statistics*, 93(2), 468–478.
- BAKER, G. P. (1992): "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100(3), 598–614.
- BAKER, G. P., AND T. N. HUBBARD (2003): "Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information," *American Economic Review*, 93(3), 551–572.

- BANDIERA, O., I. BARANKAY, AND I. RASUL (2010): "Social Incentives in the Workplace," *Review of Economic Studies*, 77(2), 417–458.
- (2013): "Team Incentives: Evidence from a Firm Level Experiment," *Journal of the European Economic Association*, 11(5), 1079–1114.
- BARANKAY, I. (2012): "Rank Incentives: Evidence from a Randomized Workplace Experiment," *Working paper*.
- BARKENBUS, J. N. (2010): "Eco-driving: An overlooked climate change initiative," *Energy Policy*, 30, 762–769.
- BARSEGHYAN, L., F. MOLINARI, T. O'DONOGHUE, AND J. C. TEITELBAUM (2018): "Estimating Risk Preferences in the Field," *Journal of Economic Literature*, 56(2), 501–564.
- BARTEL, A., C. ICHNIOWSKI, AND K. SHAW (2004): "Using "Insider Econometrics" to Study Productivity," *American Economic Review*, 94(2), 217–223.
- BASTECK, C., AND M. MANTOVANI (2018): "Cognitive ability and games of school choice," *Games and Economic Behavior*, 109, 156–183.
- BENABOU, R., AND M. PYCIA (2002): "Dynamic inconsistency and self-control: a planner-doer interpretation," *Economics Letters*, 77, 419–424.
- BENABOU, R., AND J. TIROLE (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652–1678.
- BERTRAND, M., E. DUFLO, AND S. MULLAINATHAN (2004): "How Much Should We Trust Differences-in-Differences Estimates," *The Quarterly Journal of Economics*, 119(1), 249–275.
- BLADER, S., C. GARTENBERG, AND A. PRAT (2020): "The Contingent Effect of Management Practices," *Review of Economic Studies*, 87(2), 721–749.
- BLAIS, A.-R., AND E. U. WEBER (2006): "A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations," *Judgment and Decision Making*, 1(1), 33–47.

- BLANES I VIDAL, J., AND M. NOSSOL (2011): "Tournaments Without Prizes: Evidence from Personnel Records," *Management Science*, 57(10), 1721–1736.
- BLOOM, N., B. EIFERT, A. MAHAJAN, D. MCKENZIE, AND J. ROBERTS (2013): "Does Management Matter? Evidence From India," *Quarterly Journal of Economics*, 128(1), 1–51.
- BLOOM, N., AND J. VAN REENEN (2007): "Measuring and Explaining Management Practices Across Firms and Countries," *Quarterly Journal of Economics*, 122(4), 1351–1408.
- BOOIJ, A. S., E. LEUVEN, AND H. OOSTERBEEK (2017): "Ability Peer Effects in University: Evidence from a Randomized Experiment," *Review of Economic Studies*, 84, 547–578.
- BRANDON, A., P. J. FERRARO, J. A. LIST, R. D. METCALFE, M. K. PRICE, AND F. RUNDHAMMER (2017): "Do The Effects of Social Nudges Persist? Theory and Evidence from 38 Natural Field Experiments," NBER working paper no. 23277.
- BROWN, B. W., AND C. E. LIEDHOLM (2002): "Can Web Courses Replace the Classroom in Principles of Microeconomics?," *American Economic Review*, 92(2), 444–448.
- BROWN, J. R., AND A. PREVITERO (2014): "Procrastination, Present-Biased Preferences, and Financial Behaviors," *NBER Working Paper Series*.
- BRYAN, G., D. KARLAN, AND S. NELSON (2010): "Commitment Devices," *Annual Review of Economics*, 2, 671–698.
- BRYNJOLFSSON, E., AND K. MCELHERAN (2016): "The Rapid Adoption of Data-Driven Decision-Making," *American Economic Review*, 106(5), 133–139.
- BURGER, N., G. CHARNESS, AND J. LYNHAM (2011): "Field and online experiments on self-control," *Journal of Economic Behavior & Organization*, 77(3), 393–404.

- BURKE, M. A., AND T. R. SASS (2013): "Classroom Peer Effects and Student Achievement," *Journal of Labor Economics*, 31(1), 51–82.
- CARVALHO, L. S., S. MEIER, AND S. W. WANG (2016): "Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday," *American Economic Review*, 106(2), 260–284.
- CASSAR, L., AND S. MEIER (2018): "Nonmonetary Incentives and the Implications of Work as a Source of Meaning," *Journal of Economic Perspectives*, 32(3), 215–238.
- CHAN, T. Y., J. LI, AND L. PIERCE (2014): "Learning from Peers: Knowledge Transfer and Sales Force Productivity Growth," *Marketing Science*, 33(4), 463–484.
- CHARNESS, G., U. GNEEZY, AND A. IMAS (2013): "Experimental methods: Eliciting risk preferences," *Journal of Economic Behavior & Organization*, 87, 43–51.
- CHARNESS, G., AND A. VICEISZA (2016): "Three Risk-elicitation Methods in the Field: Evidence from Rural Senegal," *Review of Behavioral Economics*, 3(2), 145–171.
- CHEN, D. L., M. SCHONGER, AND C. WICKENS (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.
- CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104(9), 2593–2632.
- CHUANG, Y., AND L. SCHECHTER (2015): "Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results," *Journal of Development Economics*, 117(C), 151–170.
- CLARK, D., D. GILL, V. PROWSE, AND M. RUSH (2017): "Using Goals to Motivate College Students: Theory and Evidence from Field Experiments," *NBER Working Paper Series*, (WP No. 23638).

- COFFMAN, L. C., M. NIEDERLE, AND A. J. WILSON (2017): "A Proposal to Organize and Promote Replications," *American Economic Review*, 107(5), 41–45.
- CORNELISSEN, T., C. DUSTMANN, AND U. SCHOENBERG (2017): "Peer Effects in the Workplace," *American Economic Review*, 107(2), 425–456.
- COX, J. C., AND G. W. HARRISON (2008): *Risk Aversion in Experiments*. Emerald Group Publishing Limited, United Kingdom.
- CROSETTO, P., AND A. FILIPPIN (2012): "The Bomb Risk Elicitation Task," *DIW SOEP Working Paper Nr. 517*.
- (2013): "The bomb risk elicitation task," *Journal of Risk and Uncertainty*, 47(1), 31–65.
- (2016): "A theoretical and experimental appraisal of four risk elicitation methods," *Experimental Economics*, 19(613–641).
- CROSON, R., AND U. GNEEZY (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47(2), 448–474.
- DAVE, C., C. C. ECKEL, C. A. JOHNSON, AND C. ROJAS (2010): "Eliciting risk preferences: When is simple better?," *Journal of Risk and Uncertainty*, 41(3), 219–243.
- DECK, C., J. LEE, J. A. REYES, AND C. C. ROSEN (2013): "A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes," *Journal of Economic Behavior & Organization*, 87, 1–24.
- DELFGAAUW, J., R. DUR, J. SOL, AND W. VERBEKE (2013): "Tournament Incentives in The Field: Gender Differences in The Workplace," *Journal of Labor Economics*, 32(2), 305–326.
- DOHMEN, T., A. FALK, K. FLIESSBACH, U. SUNDE, AND B. WEBER (2011): "Relative versus absolute income, joy of winning, and gender: Brain imaging evidence," *Journal of Public Economics*, 95(279–285).

- DOHMEN, T., A. FALK, D. HUFFMAN, U. SUNDE, J. SCHUPP, AND G. G. WAGNER (2011): "Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences," *Journal of the European Economic Association*, 9(3), 522–550.
- DUCKWORTH, A. L., E. P. SHULMAN, A. J. MASTRONARDE, S. D. PATRICK, J. ZHANG, AND J. DRUCKMAN (2015): "Will not want: Self-control rather than motivation explains the female advantage in report card grades," *Learning and Individual Differences*, 39, 13–23.
- DUNSCH, F. A., D. K. EVANS, E. EZE-AJOKU, AND M. MACIS (2017): "Management, Supervision, and Healthcare: A Field Experiment," NBER working paper no. 23749.
- ECKEL, C. C., AND P. J. GROSSMAN (2002): "Sex differences and statistical stereotyping in attitudes toward financial risk," *Evolution and Human Behavior*, 23(4), 281–295.
- EDMUNDS, A., AND A. MORRIS (2000): "The problem of information overload in business organisations: A review of the literature," *International Journal of Information Management*, 20, 17–28.
- EINAV, L., A. FINKELSTEIN, I. PASCU, AND M. R. CULLEN (2012): "How General Are Risk Preferences? Choices under Uncertainty in Different Domains," *American Economic Review*, 102(6), 2606–2638.
- ERIKSSON, T., A. POULSEN, AND M. C. VILLEVAL (2009): "Feedback and incentives: Experimental evidence," *Labour Economics*, 16(6), 679–688.
- ESCUETA, M., A. J. NICKOW, P. OREOPOULOS, AND V. QUAN (forthcoming): "Upgrading Education with Technology: Insights from Experimental Research," *Journal of Economic Literature*.
- FALK, A., AND A. ICHINO (2006): "Clean Evidence on Peer Effects," *Journal of Labor Economics*, 24(1), 39–57.
- FERRARO, P. J., AND M. K. PRICE (2013): "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment," *The Review of Economics and Statistics*, 95(1), 64–73.

- FREDERICK, S. (2003): *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*chap. Time preference and personal identity, pp. 89–114. Russell Sage Foundation, New York.
- FREEMAN, R. B. (1981): “Union Wage Practices and Wage Dispersion within Establishments,” *Industrial and Labor Relations Review*, 36(1), 3–21.
- FRUEHWIRTH, J. C. (2013): “Identifying peer achievement spillovers: Implications for desegregation and the achievement gap,” *Quantitative Economics*, 4, 85–124.
- GERARDEN, T. D., R. G. NEWELL, AND R. N. STAVINS (2017): “Assessing the Energy-Efficiency Gap,” *Journal of Economic Literature*, 55(4), 1486–1525.
- GINÉ, X., J. GOLDBERG, D. SILVERMAN, AND D. YANG (2018): “Revising Commitments: Field Evidence on the Adjustment of Prior Choices,” *The Economic Journal*, 128(608), 159–188.
- GIOIA, F. (2017): “Peer effects on risk behaviour: the importance of group identity,” *Experimental Economics*, 20(1), 100–129.
- GNEEZY, U., AND J. POTTERS (1997): “An Experiment on Risk Taking and Evaluation Periods,” *The Quarterly Journal of Economics*, 112(2), 631–645.
- GOLDFARB, A., AND C. TUCKER (2019): “Digital Economics,” *Journal of Economic Literature*, 57(1), 3–43.
- GOLMAN, R., D. HAGMANN, AND G. LOEWENSTEIN (2017): “Information Avoidance,” *Journal of Economic Literature*, 55(1), 96–135.
- GOSNELL, G. K., J. A. LIST, AND R. METCALFE (2020): “The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains,” *Journal of Political Economy*, 128(4), 1195–1233.
- GRANT, A. M. (2008): “The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions,” *Journal of Applied Psychology*, 93(1), 108–124.

- HAHN, R., R. D. METCALFE, D. NOVGORODSKY, AND M. K. PRICE (2016): "The Behavioralist as Policy Designer: The Need to Test Multiple Treatments to Meet Multiple Targets," *NBER Working Paper Series*, (No. 22886).
- HAMERMESH, D. S. (2007): "Viewpoint: Replication in Economics," *Canadian Journal of Economics*, 40(3), 715–733.
- HANNA, R., S. MULLAINATHAN, AND J. SCHWARTZSTEIN. (2014): "Learning Through Noticing: Theory and Experimental Evidence in Farming," *Quarterly Journal of Economics*, 129(3), 1311–1353.
- HANOCH, Y., J. G. JOHNSON, AND A. WILKE (2006): "Domain Specificity in Experimental Measures and Participant Recruitment: An Application to Risk-Taking Behavior," *Psychological Science*, 17(4), 300–304.
- HARRISON, G. W., M. I. LAU, AND E. E. RUTSTRÖM (2007): "Estimating Risk Attitudes in Denmark: A Field Experiment," *Scandinavian Journal of Economics*, 109(2), 341–368.
- HARRISON, G. W., AND J. A. LIST (2004): "Field Experiments," *Journal of Economic Literature*, 42(4), 1009–1055.
- HE, P. (2018): "Can self-assessed risk attitudes predict behavior under risk? Evidence from a field study in China," *Economics Letters*, 172, 107–109.
- HEFFETZ, O., T. O'DONOGHUE, AND H. S. SCHNEIDER (2016): "Forgetting and Heterogeneity in Task Delay: Evidence from New York City Parking-Ticket Recipients," *NBER Working Paper Series*, (WP No. 23012).
- HERNANDEZ-JULIAN, R., AND C. PETERS (2012): "Does the Medium Matter? Online versus Paper Coursework," *Southern Economic Journal*, 78(4), 1333–1345.
- HERSHFIELD, H. E. (2011): "Future Self-Continuity: How Conceptions of the Future Self Transform Intertemporal Choice," *Annals of the New York Academy of Sciences*, (1235), 30–43.

- HERSHFIELD, H. E., D. G. GOLDSTEIN, W. F. SHARPE, J. FOX, L. YEYKELIS, L. L. CARSTENSEN, AND J. N. BAILENSON (2011): "Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self," *Journal of Marketing Research*, 48, 23–37.
- HILLENBRAND, A., AND F. WINTER (2018): "Volunteering under population uncertainty," *Games and Economic Behavior*, 109, 65–81.
- HIMMLER, O., R. JÄCKLE, AND P. WEINSCHENK (2019): "Soft Commitments, Reminders, and Academic Performance," *American Economic Journal: Applied Economics*, 11(2), 114–142.
- HITT, L. M., AND E. BRYNJOLFSSON (1997): "Information Technology and Internal Firm Organization: An Exploratory Analysis," *Journal of Management Information Systems*, 14(2), 81–101.
- HOLLADAY, S. J., J. LARIVIERE, D. M. NOVGORODSKY, AND M. K. PRICE (2016): "Asymmetric Effects of Non-Pecuniary Signals on Search and Purchase Behavior for Energy-Efficient Durable Goods," *NBER Working Paper Series*, (No. 22939).
- HÖLMSTROM, B. (1979): "Moral Hazard and Observability," *The Bell Journal of Economics*, 10(1), 74–91.
- HÖLMSTROM, B., AND P. MILGROM (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7, 24–52.
- HOLT, C. A., AND S. K. LAURY (2002): "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), 1644–1655.
- HOLZMEISTER, F., AND A. PFURTSCHELLER (2016): "oTree: The bomb risk elicitation task," *Journal of Behavioral and Experimental Finance*, 10, 105–108.
- HOXBY, C. M., AND G. WEINGARTH (2005): "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects," .
- HSIAW, A. (2015): "Goal Bracketing and Self-Control," *Working paper*.

- HUBBARD, T. N. (2003): "Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking," *American Economic Review*, 93(4), 1328–1353.
- ICCT (2013): "From Laboratory to Road: A Comparison of Official and 'Real-World' Fuel Consumption CO₂ Values for Cars in Europe and United States," Discussion paper, International Council on Clean Transportation.
- IEA (2012): "Improving the Fuel Economy of Road Vehicles: A Policy Package," Discussion paper, International Energy Agency.
- JACKSON, C. K., AND E. BRUEGMANN (2009): "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics*, 1(4), 85–108.
- JACKSON, C. K., AND H. S. SCHNEIDER (2015): "Checklists and Worker Behavior: A Field Experiment," *American Economic Journal: Applied Economics*, 7(4), 136–168.
- JACOBSON, S., AND R. PETRIE (2009): "Learning from mistakes: What do inconsistent choices over risk tell us?," *Journal of Risk and Uncertainty*, 38(2), 143–158.
- JAEGER, D. A., T. DOHMEN, A. FALK, D. HUFFMAN, U. SUNDE, AND H. BONIN (2010): "Direct Evidence on Risk Attitudes and Migration," *Review of Economics and Statistics*, 92(3), 684–689.
- KAIVANTO, K., AND E. B. KROLL (2012): "Alternation Bias and Reduction in St. Petersburg Gambles: An Experimental Investigation," *Working paper*.
- KAUR, S., M. KREMER, AND S. MULLAINATHAN (2010): "Self-Control and the Development of Work Arrangements," *American Economic Review*, 100(2), 624–628.
- KELLEY, E. M., G. LANE, AND D. SCHÖNHOLZER (2018): "The Impact of Monitoring Technologies on Contracts and Employee Behavior: Ex-

- perimental Evidence from Kenya's Transit Industry," mimeo, University of Berkeley.
- KLUGER, A. N., AND A. DENISI (1996): "The Effects of Information Interventions on Performance: A Historical Review, a Meta-Analysis and a Preliminary Information Intervention Theory," *Psychological Bulletin*, 119(2), 254–284.
- KOCH, A., J. NAFZIGER, AND H. S. NIELSEN (2015): "Behavioral economics of education," *Journal of Economic Behavior & Organization*, 115, 3–17.
- KOCH, A. K., AND J. NAFZIGER (2016): "Goals and bracketing under mental accounting," *Journal of Economic Theory*, 162, 305–351.
- (2017): "Motivational Goal Bracketing: An Experiment," *Working paper*.
- KOUDSTAAL, M., R. SLOOF, AND M. VAN PRAAG (2016): "Risk, Uncertainty, and Entrepreneurship: Evidence from a Lab-in-the-Field Experiment," *Management Science*, 62(10), 2897–2915.
- KUHNEN, C. M., AND A. TYMULA (2012): "Feedback, Self-Esteem, and Performance in Organizations," *Management Science*, 58(1), 94–113.
- LAIBSON, D. (1997): "Golden Eggs and Hyperbolic Discounting," *The Quarterly Journal of Economics*, 112(2), 443–478.
- LAM, C. F., D. S. DERUE, E. P. KARAM, AND J. R. HOLLENBECK (2011): "The impact of feedback frequency on learning and task performance: Challenging the "more is better" assumption," *Organizational Behavior and Human Decision Processes*, 116(2), 217–228.
- LARRICK, R. P., AND J. B. SOLL (2008): "The MPG Illusion," *Science*, 320(5883), 1593–1594.
- LAVECCHIA, A. M., H. LIU, AND P. OREOPOULOS (2016): "Behavioral economics of education: progress and possibilities," *Handbook of the Economics of Education*, 5, 1–74.

- LAVY, V., M. D. PASERMAN, AND A. SCHLOSSER (2011): "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom," *The Economic Journal*, 122, 208–237.
- LAVY, V., O. SILVA, AND F. WEINHARDT (2012): "The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools," *Journal of Labor Economics*, 30(2), 367–414.
- LAZEAR, E. P., AND S. ROSEN (1981): "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89(5), 841–864.
- LEE, W., R. H. COURTNEY, AND S. J. BALASSI (2010): "Do Online Homework Tools Improve Student Results in Principles of Microeconomics Courses?," *American Economic Review*, 100(2), 283–286.
- LEJUEZ, C. W., J. P. READ, C. W. KAHLER, J. B. RICHARDS, S. E. RAMSEY, G. L. STUART, D. R. STRONG, AND R. A. BROWN (2002): "Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART)," *Journal of Experimental Psychology: Applied*, 8(2), 75–84.
- LEUVEN, E., H. OOSTERBEEK, AND B. VAN DER KLAUW (2010): "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment," *Journal of the European Economic Association*, 8(6), 1243–1265.
- LEVITT, S. D., J. A. LIST, S. NECKERMANN, AND S. SADOFF (2016): "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, 8(4), 183–219.
- LINDQUIST, M., J. SAUERMAN, AND Y. ZENOU (2017): "Peer Effects in the Workplace: A Network Approach," *CEPR Discussion Paper 10928*.

- LIST, J. A. (2003): "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy*, 2006(1), 1–37.
- LIST, J. A., A. M. SHAIKH, AND Y. XU (2019): "Multiple hypothesis testing in experimental economics," *Experimental Economics*, 22, 773–793.
- MARTINEZ, S.-K., S. MEIER, AND C. SPRENGER (2017): "Procrastination in the Field: Evidence from Tax Filing," *Working paper*.
- MAS, A., AND E. MORETTI (2009): "Peers at Work," *American Economic Review*, 99(1), 112–45.
- MEIER, S., AND C. D. SPRENGER (2015): "Temporal Stability of Time Preferences," *The Review of Economics and Statistics*, 97(2), 273–286.
- METCALFE, R., S. BURGESS, AND S. PROUD (2019): "Students' effort and educational achievement: Using the timing of the World Cup to vary the value of leisure," *Journal of Public Economics*, 172, 1111–1126.
- MIVW (2010): "Public Transport in the Netherlands," Discussion paper, Ministry of Transport, Public Works and Water Management.
- MOLDOVANU, B., A. SELA, AND X. SHI (2007): "Contests for Status," *Journal of Political Economy*, 115(2), 338–363.
- MUELLER-LANGER, F., B. FECHER, D. HARHOFF, AND G. G. WAGNER (2019): "Replication studies in economics—How many and which papers are chosen for replication, and why?," *Research Policy*, 48(1), 62–83.
- NILEKANI, J. (2018): "Driving Down Demand for Diesel: Does a Bus Driver Training and Incentive Program Increase Fuel Efficiency?," working paper.
- O'DONOGHUE, T., AND M. RABIN (1999): "Doing It Now or Later," *American Economic Review*, 89(1), 103–124.
- OECD (2008): "Education at a Glance 2008: OECD Indicators," Discussion paper, OECD.

- OREOPOULOS, P., R. W. PATTERSON, U. PETRONIJEVIC, AND N. G. POPE (2018): "Lack of Study Time is the Problem, but What is the Solution? Unsuccessful Attempts to Help Traditional and Online College Students," *NBER Working Paper Series*, (WP No. 25036).
- OREOPOULOS, P., AND U. PETRONIJEVIC (2018): "Student Coaching: How Far Can Technology Go?," *Journal of Human Resources*, 53(2), 299–329.
- PAOLA, M. D., AND V. SCOPPA (2015): "Procrastination, academic success and the effectiveness of a remedial program," *Journal of Economic Behavior & Organization*, 115, 217–236.
- PAPAY, J. P., E. S. TAYLOR, J. H. TYLER, AND M. E. LASKI (2020): "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data," *American Economic Journal: Economic Policy*, 12(1), 359–388.
- PIERCE, L., D. C. SNOW, AND A. MCAFEE (2015): "Cleaning House: The Impact of Information Technology Monitoring on Employee Theft and Productivity," *Management Science*, 61(10), 2299–2319.
- REUBEN, E., P. SAPIENZA, AND L. ZINGALES (2015): "Procrastination and Impatience," *Journal of Behavioral and Experimental Economics*, 58, 63–76.
- ROMENSEN, G.-J. (2020a): "Learning From High-Achieving Peers: A Quasi-Experiment in the Field," *Working paper*.
- (2020b): "Rabbits and Study Habits: A Field Experiment on Pacesetters and Student Effort," *Working paper*.
- ROMENSEN, G.-J., AND A. R. SOETEVENT (2020a): "Improving Worker Productivity Through Tailored Performance Feedback: Field Experimental Evidence from Bus Drivers," *Working paper*.
- (2020b): "P(l)aying at the Pump: Incentivized Measurement of Risk Attitudes in the Field," *Working paper*.

- RYAN, R. M., AND E. L. DECI (2000): "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist*, 55(1), 68–78.
- SACERDOTE, B. (2011): "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?," in *Handbook of the Economics of Education*, ed. by L. W. Eric A. Hanushek, Stephen Machin, vol. 3, chap. 4, pp. 249–277. Elsevier.
- SANDVIK, J., R. SAOUMA, N. SEEGER, AND C. STANTON (2020): "Workplace Knowledge Flows," CEPR Discussion Paper No. DP14299.
- SCHILDBERG-HÖRISCH, H. (2018): "Are Risk Preferences Stable?," *Journal of Economic Perspectives*, 32(2), 135–154.
- SIMON, H. A. (1973): "Applying Information Technology to Organization Design," *Public Administration Review*, 33(3), 268–278.
- SKRIABIKOVA, O. J., T. DOHMEN, AND B. KRIECHEL (2014): "New evidence on the relationship between risk attitudes and self-employment," *Labour Economics*, 30, 176–184.
- SONG, H., A. L. TUCKER, K. L. MURRELL, AND D. R. VINSON (2018): "Closing the Productivity Gap: Improving Worker Productivity Through Public Relative Performance Feedback and Validation of Best Practices," *Management Science*, 64(6), 2628–2649.
- SOSIN, K., B. J. LECHA, R. AGARWAL, R. L. BARTLETT, AND J. I. DANIEL (2004): "Efficiency in the Use of Technology in Economic Education: Some Preliminary Results," *American Economic Review*, 94(2), 253–258.
- STAATS, B. R., H. DAI, D. HOFMANN, AND K. L. MILKMAN (2017): "Motivating Process Compliance Through Individual Electronic Monitoring: An Empirical Examination of Hand Hygiene in Healthcare," *Management Science*, 63(5), 1563–1585.

- STEEL, P. (2007): "The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure," *Psychological Bulletin*, 133(1), 65–94.
- STINEBRICKNER, R., AND T. R. STINEBRICKNER (2008): "The Causal Effect of Studying on Academic Performance," *The B.E. Journal of Economic Analysis & Policy*, 8(1), 1–55.
- SWOBODA, A., AND L. FEILER (2016): "Measuring the Effect of Blended Learning: Evidence from a Selective Liberal Arts College," *American Economic Review*, 106(5), 368–372.
- SYVERSON, C. (2011): "What Determines Productivity?," *Journal of Economic Literature*, 49(2), 326–365.
- THALER, R. H., AND H. M. SHEFRIN (1981): "An Economic Theory of Self-Control," *Journal of Political Economy*, 89(2), 392–406.
- TRAN, A., AND R. ZECKHAUSER (2012): "Rank as an inherent incentive: Evidence from a field experiment," *Journal of Public Economics*, 96(9–10), 645–650.
- VON GAUDECKER, H.-M., A. VAN SOEST, AND E. WENGSTROM (2011): "Heterogeneity in Risky Choice Behavior in a Broad Population," *American Economic Review*, 101(2), 664–694.
- WALDINGER, F. (2012): "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany," *Review of Economic Studies*, 79(2), 838–861.
- WEIMANN, J., AND J. BROSIG-KOCH (2019): *Methods in Experimental Economics*. Springer International Publishing.
- WONG, W.-K. (2008): "How much time-inconsistency is there and does it matter? Evidence on self-awareness, size, and effects," *Journal of Economic Behavior & Organization*, 68(3–4), 645–656.

Chapter 6

Samenvatting (Dutch summary)

De digitalisering van economische en sociale activiteiten behoort tot de zichtbaarste ontwikkelingen van onze tijd. Individuele keuzes en acties worden door digitale technologieën minutieus vastgelegd en opgeslagen tegen lagere kosten dan ooit tevoren. In verscheidene principaal-agent relaties biedt dit mogelijkheden om nauwkeurig inzicht te krijgen in bijvoorbeeld de determinanten van werknemersproductiviteit, consumentenkeuzes en studie-inzet van studenten. De invulling en evaluatie van feedback kan hierdoor in potentie met aanzienlijk meer oog voor maatwerk worden vormgegeven. Een open vraag is echter hoe feedback op maat het best geleverd kan worden in een gedigitaliseerde omgeving waarin de principaal volledig zicht heeft op de inzet en prestaties van de agent.

Dit proefschrift richt zich op feedbackontwerp en elicitatie van individuele preferenties in gedigitaliseerde omgevingen. Het proefschrift bestaat uit vier veldexperimenten. De eerste twee experimenten zijn gericht op feedbackprovisie aan werknemers, waarbij gebruik wordt gemaakt van een recent geïntroduceerde monitoring-technologie bij een veldpartner. Het doel is te onderzoeken welke vormen van gepersonaliseerde feedback effectief zijn in het verbeteren van de werknemersproductiviteit. Het derde experiment maakt gebruik van een online leerplatform om de studie-inzet van studenten te meten en om individuele feedback te communiceren die de inzet en leeruitkomsten kan verbeteren. In het laatste experiment worden individuele risicovoorkeuren van consumenten gemeten middels twee elicitatie-

methoden die via een communicatietechnologie in *real time* worden gepresenteerd aan de consumenten. Hieronder volgt een beknopte samenvatting van deze vier studies.

In hoofdstuk 2 en hoofdstuk 3 wordt samengewerkt met een vervoersorganisatie om te onderzoeken hoe digitale technologieën gebruikt kunnen worden in het ontwerpen en evalueren van geïndividualiseerde feedback voor werknemers. Gedurende de studieperiode heeft de organisatie een elektronisch boordcomputersysteem geïnstalleerd in de busvloot van een aantal concessiegebieden in Nederland. Dit systeem meet minutieus de prestaties van buschauffeurs op verschillende rijdimensies en geeft daarmee een gedetailleerd beeld van productiviteit door de tijd heen. Naast brandstofverbruik bevat het systeem ook informatie over accelereren, remmen en het nemen van bochten. De organisatie heeft daarmee toegang tot zowel een geaggregeerde maatstaf van werknemersproductiviteit als gedesaggregeerde maatstaven.

In hoofdstuk 2 wordt het elektronisch boordcomputersysteem gebruikt om een veldexperiment te implementeren onder 409 buschauffeurs. Het doel van het experiment is het testen van twee veelbelovende vormen van feedback op maat: een rapport waarin de productiviteit wordt vergeleken met die van collega's (*peers*) en een programma waarin chauffeurs deelnemen aan een individuele coachingsessie met een ervaren collega. In bestaand onderzoek over peer-vergelijkingen is feedback vaak op basis van het eindresultaat in plaats van op basis van de tussenstappen die leiden tot dat resultaat. Een mogelijk gevolg is dat de ontvangers van feedback weinig tot geen sturing krijgen op welke gebieden zij zich kunnen verbeteren om een beter eindresultaat te bewerkstelligen. Dit kan een demotiverend effect hebben op de ontvangers.

De gedachtegang achter het veldexperiment is dat werknemers vaak meer controle hebben over tussenstappen dan over einduitkomsten, net zoals bij studenten die zelf beslissen hoeveel uur ze willen studeren voor een tentamen (zie ook hoofdstuk 4 in dit proefschrift). Een hogere mate van controle kan een positief effect hebben op de motivatie en zo de werknemersproductiviteit verhogen. Het experiment is als volgt opgebouwd. Allereerst worden er peer-groepen gemaakt met chauffeurs van dezelfde vestiging.

Vervolgens wordt de maandelijkse relatieve prestatie van een chauffeur in de groep gebruikt om berichten met peer-vergelijkingen op te stellen, waarmee een chauffeur inzicht krijgt in of hij of zij het goed of slecht heeft gedaan op één van de gedesaggregeerde rijdimensies (accelereren, remmen en bochten nemen). De berichten worden geïntegreerd in maandelijkse feedback rapporten die gedistribueerd worden onder de chauffeurs. De feedback in de berichten is geïndividualiseerd op basis van experimentele variatie in de benadering (positief of negatief) en compositie (alleen negatief of een mix van positief en negatief). Op deze manier kan worden onderzocht welke vorm van feedback op maat het beste werkt in het verbeteren van de werknemersproductiviteit.

De analyse in het veldexperiment maakt gebruik van meer dan 500.000 ritobservaties die zijn verzameld gedurende een studieperiode van twee jaar. De belangrijkste bevinding is dat de berichten met peer-vergelijkingen op verschillende rijdimensies in het algemeen geen effect lijken te hebben op zowel brandstofverbruik als de rijdimensies. De prestaties van chauffeurs in de interventiegroepen verschillen niet significant met die van chauffeurs in een controlegroep waarin alleen generieke feedback wordt gegeven. Alleen chauffeurs die nog niet hebben deelgenomen aan een individuele coachingsessie met een collega lijken enigszins het rijgedrag te verbeteren nadat zij correctieve berichten hebben ontvangen.

Individuele coaching door een ervaren collega (*peer coaching*) is een andere veelbelovende manier om feedback op maat te geven in werkgerelateerde omgevingen. Bestaand onderzoek stelt dat coaching kennisoverdracht en de uitwisseling van vaardigheden kan faciliteren. Of dit ook leidt tot een verbetering in de productiviteit van werknemers is vooralsnog een open vraag in de literatuur. Om dit te onderzoeken wordt een coachingprogramma van de vervoersorganisatie geëvalueerd dat parallel loopt aan het feedbackprogramma met de peer-vergelijkingen. In het programma rijdt een ervaren collega als coach onaangekondigd mee met een buschauffeur gedurende een deel van de dienst. Na afloop van een rit geeft de coach feedback op maat op basis van observaties tijdens de rit (coaches hebben geen toegang tot de data van het boordcomputersysteem).

In hoofdstuk 2 wordt er gebruikgemaakt van het feit dat de coaches

in het concessiegebied van de studie een dagboek hebben bijgehouden van hun coaching-activiteiten. Deze dagboeken maken het mogelijk om voor alle chauffeurs vast te stellen wanneer ze hebben deelgenomen aan een coachingsessie. Middels een quasi-experimentele evaluatie kan vervolgens worden vastgesteld wat het effect is van deelname aan een coachingsessie op de werknemersproductiviteit. De resultaten van deze evaluatie laten zien dat peer coaching een positief effect heeft op alle rijdimensies. De positieve effecten zijn voornamelijk zichtbaar bij chauffeurs die aanvankelijk minder presteerden. De verbeteringen houden een aantal weken aan, waarna het effect lijkt weg te ebben. Dit suggereert dat chauffeurs de feedback van de coach uiteindelijk niet hebben geïnternaliseerd. Additionele interventies lijken daarom nodig te zijn om de waargenomen verbeteringen structureel te maken.

Een beperking van het onderzoek in hoofdstuk 2 is dat twee parallelle feedbackprogramma's tegelijkertijd worden geïmplementeerd en geëvalueerd. Er kunnen hierdoor interacties tussen de programma's ontstaan die mogelijk niet volledig worden geïdentificeerd in de gepresenteerde analyses. Om uit te sluiten dat dergelijke interacties een invloed hebben gehad op de resultaten van het coachingprogramma wordt in hoofdstuk 3 een replicatie uitgevoerd. In de replicatie wordt een vergelijkbaar coachingprogramma geëvalueerd in twee andere concessiegebieden van de vervoersorganisatie. In beide gebieden loopt geen parallel feedbackprogramma met peer-vergelijkingen. Dagboeken van coaches worden opnieuw gebruikt om per chauffeur vast te stellen wanneer zij zijn gecoacht. De analyse in deze replicatie maakt gebruik van ongeveer twee miljoen ritobservaties van 1.032 buschauffeurs. De resultaten van de replicatie zijn in lijn met de resultaten van het onderzoek in hoofdstuk 2. Buschauffeurs presteren voor een aantal weken beter op meerdere rijdimensies nadat zij zijn gecoacht door een ervaren collega. De verbeteringen in accelereren en remmen lijken meer structureel te zijn dan in hoofdstuk 2.

Bestaande studies naar peer-effecten in werkgerelateerde omgevingen documenteren vaak dat peer-effecten ontstaan vanwege sociale druk. In deze studies voeren werknemers relatief simpele taken uit met weinig leermogelijkheden. Daarnaast kunnen werknemers eenvoudig de productiviteit

van collega's observeren. Echter, in moderne werkomgevingen, zoals de transportsector in hoofdstuk 2 en hoofdstuk 3, zijn dit soort eigenschappen afwezig. Sociale druk kan ten dele verklaren waarom chauffeurs beter presteren tijdens de coachingsessie, maar het kan niet verklaren waarom na coaching verbeteringen worden geobserveerd. De resultaten zijn daarom meer in lijn met de interpretatie dat werknemers (tijdelijk) nieuwe kennis en vaardigheden vergaren doordat ze leren van collega's.

In hoofdstuk 4 wordt een nieuw online leerplatform ontwikkeld en gebruikt om een veldexperiment te implementeren onder 573 studenten. Het experiment richt zich op uitstelgedrag onder studenten. In bestaand onderzoek is aangetoond dat veel studenten beneden hun academisch niveau presteren vanwege een gebrek aan studie-inzet. Recente studies suggereren ook dat een hogere studie-inzet het best bereikt kan worden door interventies te richten op inzet in plaats van op einduitkomsten. Op dit vlak is tot dusver echter weinig vooruitgang geboekt vanwege een gebrek aan informatie over de studie-inzet en studieplannen van studenten. Uitstelgedrag onder studenten en een gebrek aan zelfcontrole vergroten de uitdaging om studenten meer te laten studeren.

In het veldexperiment formuleren studenten elke week een studiedoel en studieplan voor de week. Deze informatie wordt vervolgens gebruikt om geïndividualiseerde gangmakers (*hazen*) te construeren. Een gangmaker is een bewegend referentiepunt dat beweegt op het tempo van het door de student geformuleerde studieplan. De gangmaker is daarmee vergelijkbaar met hazen in bijvoorbeeld marathons. Het niet kunnen bijhouden van de gangmaker confronteert de student onmiddellijk met zijn of haar uitstelgedrag. Studenten met referentie-afhankelijke voorkeuren proberen een achterstand tot de gangmaker te vermijden vanwege verliesaversie en zullen daarom in theorie meer studie-inzet tonen.

Het veldexperiment heeft als doel het causale effect van de gangmaker op de studie-inzet en leeruitkomsten van studenten te identificeren. De experimentele opzet maakt het mogelijk om te controleren voor de effecten die voortvloeien uit het formuleren van studiedoelen en studieplannen. Om studie-inzet te meten en om de gangmaker in *real time* te presenteren aan de studenten wordt gebruikgemaakt van een zelfontwikkeld online leer-

platform. Het platform wordt gratis aangeboden aan studenten als een niet-verplicht onderdeel in een grote universitaire cursus. Het platform bevat zelfontwikkeld oefenmateriaal, feedback en samenvattingen van cursusonderwerpen. Naast gebruiksgedata van het online leerplatform wordt in het experiment ook gebruikgemaakt van wekelijkse enquêtes over tijdsinvesteringen in bijvoorbeeld het bezoeken van colleges en het lezen van het studieboek. Op deze manier kan goed in kaart worden gebracht hoe studenten studeren door de tijd heen. Administratieve data met leeruitkomsten en achtergrondkarakteristieken van de studenten complementeren deze data in de analyse.

De bevindingen in het veldexperiment zijn als volgt. Allereerst laat het experiment zien dat studenten ambitieuzere studiedoelen formuleren wanneer zij een gangmaker kunnen observeren die beweegt op het tempo van hun eigen studieplan. De studiedoelen zijn ongeveer twee keer zo hoog als in de controlegroep. De ambitieuzere studiedoelen vertalen zich echter niet in meer studie-inzet, zoals gemeten op het online leerplatform en in de wekelijkse enquêtes. Een verdere analyse toont aan dat studenten in het begin van de week al achterlopen op hun gangmaker en gemiddeld genomen geen inhaalslag maken later in de week. Ook is er geen empirisch bewijs dat de gangmaker een effect heeft op de leeruitkomsten van de studenten. Tenslotte is er weinig tot geen bewijs van heterogeniteit in de resultaten op basis van geslacht, leervermogen en neigingen tot uitstelgedrag.

In hoofdstuk 5 wordt een samenwerking gestart met een veldpartner die een aantal onbemande tankstations exploiteert in Nederland. De tankstations zijn voorzien van innovatieve communicatietechnologieën die directe communicatie met de klant mogelijk maken. Tankstations lenen zich goed als onderzoekssetting omdat het een natuurlijke omgeving is waar een brede dwarsdoorsnede van de bevolking boven de 18 jaar te vinden is. In hoofdstuk 5 wordt deze context gebruikt om te onderzoeken hoe digitale technologieën in het veld gebruikt kunnen worden om de risicovoorkeuren van een grote en diverse consumentengroep te meten.

Risicovoorkeuren vervullen een centrale rol bij keuzes die gemaakt worden in onzekerheid. Onderzoek naar hoe individuele risicovoorkeuren het best gemeten kunnen worden is daarom belangrijk. Uit de bestaande lit-

eratuur blijkt dat het vaak praktisch niet haalbaar is om in het veld de risicovoorkeuren van een grote groep consumenten te meten op een manier die *incentive*-compatibel is. Dit heeft ertoe geleid dat er grofweg twee populaire methodes zijn om individuele risicovoorkeuren te meten. De populairste methode in het veld is om respondenten via een enquête naar hun risicobereidheid te vragen. Deze methode is snel en goedkoop, maar in veel gevallen niet *incentive*-compatibel. In lab-studies worden risicovoorkeuren vaak gemeten middels een experimentele methode die nauwkeurig en *incentive*-compatibel is, maar tegelijkertijd meer tijd vergt en hogere implementatiekosten kent. In hoeverre beide methodes in het veld tot vergelijkbare schattingen van risicovoorkeuren komen is vooralsnog niet uitgebreid onderzocht.

In hoofdstuk 5 wordt de digitale infrastructuur van 13 geografisch verspreide tankstations in Nederland gebruikt om beide methodes meermaals voor te leggen aan consumenten. Eens per maand wordt een consument na het tanken via het scherm op de pomp uitgenodigd om de *Bomb Risk Elicitation Task* (BRET) te spelen en om de vraag over risicobereidheid van de Duitse *Socio-Economic Panel Study* (SOEP) te beantwoorden. In de BRET, geïntroduceerd en gevalideerd door Crosetto en Filippin (2013), worden consumenten geconfronteerd met 100 dozen, waarvan één willekeurige doos een bom bevat. De verdiensten in het spel nemen lineair toe met het aantal verzamelde dozen, maar vallen terug tot nul wanneer de doos met de bom onderdeel is van de verzamelde dozen. In de SOEP vraag, gevalideerd door Dohmen et al. (2011), worden consumenten gevraagd om op een schaal van nul (niet bereid) tot tien (zeer bereid) aan te geven hoe risicobereid ze zichzelf schatten.

In de studieperiode van één jaar hebben ongeveer 2.500 consumenten beide methodes tenminste één keer voltooid op de pomp. Uit de antwoorden blijkt dat de meerderheid van de consumenten in deze natuurlijke omgeving, ongeveer 82 procent, risico-averse keuzes maakt in the BRET. Er zijn geen verschillen zichtbaar in de BRET keuzes tussen mannen en vrouwen. Ook demografische kenmerken, zoals leeftijd en inkomen, hebben geen invloed op het aantal verzamelde dozen in de BRET. Het meest gegeven antwoord op de SOEP vraag is zeven. De zelfbeoordeelde risicobereidheid onder

consumenten is dus relatief hoog, en dit lijkt met name te gelden voor mannen. Een verdere analyse laat zien dat er een positieve maar kleine correlatie is tussen keuzes in de BRET en antwoorden op de SOEP. Tijdens herhalingen van beide methodes is er in het algemeen veel variatie in de keuzes en antwoorden van de consumenten.

De studies in dit proefschrift laten de vele mogelijkheden en enkele beperkingen zien van digitale technologieën voor feedbackontwerp en elicitatie van individuele preferenties. Op basis van de gepresenteerde resultaten zijn er tenminste drie veelbelovende richtingen voor verder onderzoek. Allereerst hebben de studies laten zien dat digitale technologieën de mogelijkheid creëren om feedback op maat te ontwikkelen dat tijdig en gericht kan worden gedeeld met de doelgroep. Welke vorm van maatwerk het beste werkt in het verbeteren van individuele prestaties blijft een belangrijke onderzoeksvraag, in het bijzonder in omgevingen waarin financiële prikkels als ongewenst worden ervaren. Ten tweede maakt technologie het mogelijk om individuele preferenties, zoals risicovoorkeuren, direct te meten in de omgeving waarin consumentenkeuzes worden gemaakt. Dit maakt het eenvoudiger voor onderzoekers om de betreffende consumentenkeuzes te onderzoeken in situaties waarin individuele preferenties door de tijd heen kunnen variëren. Tenslotte is meer onderzoek nodig naar de vraag wanneer mensen en technologie elkaar complementeren of substitueren. Digitale technologie is goed in staat om individuele verbeterpunten te identificeren, maar is mogelijk minder effectief dan mensen in het overbrengen van de feedback dat daaruit voortvloeit.