Tailored Feedback and Worker Green Behavior: Field Evidence from Bus Drivers*

Gert-Jan Romensen†
University of Groningen

Adriaan R. Soetevent‡
University of Groningen
Tinbergen Institute

July 21, 2017

First version: 21 July 2017

Abstract

How to engage workers in conservation efforts when the company pays the bill? In a field experiment with 409 bus drivers, we investigate the potential of targeted peer-comparison feedback and on-the-road coaching. Drivers receive individualized reports with peer-comparison messages on multiple driving dimensions. In addition, coaches quasi randomly provide drivers with in person coaching moments on the bus. Based on 800,000 trip-level observations, we find that the targeted peer-comparison treatments do not improve driving. On-the-road coaching significantly improves driving on multiple dimensions but only temporarily. Further analysis reveals negative interaction effects between the two programs.

JEL classification: D2, M5, Q5.
Keywords: peer comparisons, coaching, worker motivation, fuel conservation.

*We thank Robert Durr, Marco Haan, Noemi Pace, Noémi Péter and seminar participants at M-BEES & M-BEPS 2017, IMEBESS 2017, AFE 2016, Universitat Autònoma de Barcelona, and Ca' Foscari University of Venice for their valuable comments. We are especially grateful to Peter Boersma, Marten Feenstra and Wouter van der Meer of Arriva for their time, support and excellent assistance in enabling this project. Views and opinions expressed in this paper as well as all remaining errors are solely those of the authors.

†Corresponding author. University of Groningen, Faculty of Economics and Business, Nettelbosje 2, 9747 AE Groningen, The Netherlands, g.j.romensen@rug.nl.

‡University of Groningen, Faculty of Economics and Business, Nettelbosje 2, 9747 AE Groningen, The Netherlands, a.r.soetevent@rug.nl.

1 Introduction

Transport is central to our everyday lives but 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). The largest share, about 79%, is consumed by on-the-road vehicles (EIA 2015). Fuel-efficient driving is hailed as low-hanging fruit to improve conservation levels (ICC 2013) but picking this fruit can be challenging when drivers have no financial stake in fuel savings. This concern is particularly acute within firms, where the design of conservation incentives is often dictated by institutional constraints that hinder the use of pay-for-performance schemes.[1] How to non-financially motivate workers to engage in green behaviors such as efficient driving is therefore a critical question to ask in making firms more sustainable.

Innovations in the transport sector related to on-board monitoring open up novel opportunities for tailoring and evaluating non-financial incentives.[2] Electronic on-board recorders (EOBR) enable high-frequency granular measurement of worker-level fuel consumption and allow firms to not only identify high users but also to tailor feedback by decomposing consumption into its underlying sources. This creates distinct possibilities for new data-driven designs of conservation incentives (Brynjolfsson and McElheran 2016). The link between fuel usage and production costs suggests that these incentives could create a win-win scenario for the environment and the firm (Gosnell, List and Metcalfe 2016).

In this paper, we therefore join the installation process of new EOBRs in the entire bus fleet of a large public transport company. Gathering over 800,000 trip-level observations on driving behavior, we evaluate two forms of tailored feedback among 409 bus drivers: targeted peer-comparison feedback and on-the-road coaching. In a field experiment, drivers are assessed on multiple driving dimensions and randomly assigned to individualized reports with varying numbers of peer-comparison messages. In addition, we evaluate the effects of a parallel coaching program with a quasi-experimental design. In this program, designated experienced drivers offer tailored feedback while riding along with their colleagues. Evidence shows that the selection of which driver receives in person

[1] 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 single wage rates (rather than pay based on merit).

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

coaching when can be considered random. We follow drivers for two years in order to establish a long baseline and experimental period, making it possible to measure both immediate and delayed responses to the feedback programs.

Existing studies on non-financial incentive schemes in the residential sector stress the importance of feedback and social approval (Allcott and Mullainathan 2010). For example, incorporating social comparisons in feedback reports reduces household consumption of energy (Allcott 2011) 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.[3]

A small but emerging literature considers the workplace for evidence on the effect of non-financial incentives on conservation efforts. One study reports fuel savings when airline captains are provided with tailored feedback (Gosnell et al. 2016). Other work finds that the provision of social-comparison information in plants with (without) a teamwork culture leads to decreased (improved) truck driver performance (Blader, Gartenberg and Prat 2016). This suggests that firms may fruitfully customize relative performance feedback by tailoring the content or by targeting subsets of workers (Kuhnen and Tymula 2012). Given that firms increasingly record and store data on multiple dimensions of worker-level productivity, this seems a viable and promising approach to designing conservation incentives. This raises the key issue of what level of feedback is optimal. Should managers reveal all relative positions on productivity dimensions to the worker or should the information be dosed in such a way that only selected dimensions are shown?

Previous studies indicate that relative performance feedback can improve worker productivity (Blanes i Vidal and Nossol 2011), sales growth (Delfgaauw, Dur, Sol and Verbeke 2013) and (high school) student performance (Tran and Zeckhauser 2012, Azmat and Iriberri 2010). Some studies however report decreased performance after rankings are provided (Ashraf, Bandiera and Lee 2014) and improved performance when they are

[3]In an overview article on the energy-efficiency gap, Allcott and Greenstone (2012) also call for more empirical evaluations of the impact of energy efficiency programs on heterogeneous consumer types.

abolished (Barankay 2012). Agents may exhibit rank incentives (Barankay 2012, Tran and Zeckhauser 2012) in which 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 et al. 2014).

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 and makes it easier for workers to change their behavior. Our research design does exactly that.

In close cooperation with our field partner Arriva Netherlands, we have the unique opportunity to construct rankings on the driving dimensions Acceleration, Braking, and Cornering (the so-called ABC-dimensions), and to experimentally vary the number of relative positions that will be communicated to drivers in a monthly feedback report. A driver's ABC scores co-determine more aggregate measures of productivity such as fuel efficiency and passenger comfort. In the first condition, we give drivers information on the poor relative position on one dimension only, even if the driver performs relatively poorly on multiple dimensions. That is, we deliberately withhold some rankings to allow drivers to focus their effort. The second condition 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. For example, a driver who performs poorly on acceleration but well on braking will receive a negative notification on the former dimension and a positive comment on the latter. This allows us to assess the value of providing a mix of corrective and positive feedback. In the final condition, all relative positions on driving behaviors are communicated whenever the driver performs poorly compared to a reference group of peers. Drivers in the control condition do not receive this type of personalized peer-comparison feedback. Together these interventions enable us to explore the potential of on-board monitoring technologies in the customization of relative performance feedback.

Tailoring feedback to disaggregate productivity measures can be a cost-effective means

to improve conservation efforts among workers (Gosnell et al. 2016). 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.[4] The feeling of being in control is a key source of human motivation (Ryan and Deci 2000). Identifying and targeting behaviors that contribute to environmental problems should therefore be central in the design of energy conservation interventions (Abrahamse, Steg, Vlek and Rothengatter 2005).

In the context of relative performance feedback, however, the dosage of rankings may matter as it could potentially aggravate adverse effects by making poor performance even more salient to the worker. When information directly enters the utility function (Golman, Hagmann and Loewenstein 2017), informing workers about poor performance on multiple dimensions may motivation. Dohmen et al. (2011), for example, show that reward-related brain areas negatively correlate with lower relative incomes. The increased level of detail in feedback provision may also have adverse effects similar to the finding that increasing feedback frequency could lead to more mistakes being made (Eriksson, Poulsen and Villeval 2009) and reduced task effort due to overwhelmed cognitive resources (Lam, DeRue, Karam and Hollenbeck 2011). This poses a challenge, given that poor performers have the biggest room for improvement and are thus precisely the group that one wants to target with detailed feedback. Our research design aims to address this challenge.

In our evaluation of the tailored peer-comparison treatments, we find that the aggregate performance of treated drivers is not significantly different from drivers in a control group with general feedback. The point estimates across treatments for fuel economy (km/liters), an aggregate measure of driving behavior, are small in size, about 0.007 (0.015σ)[5], and are (jointly) not statistically significant. Similar qualitative results are obtained for acceleration, braking, and cornering. Notifying treated bus drivers that peer-comparison messages are no longer included in the feedback reports does not lead to a change in driving behavior in the post-experimental period.

On-the-road coaching is more effective in improving driving behavior. The largest improvements are observed during the coaching day: fuel economy goes up (0.23σ) and drivers engage in less excessive acceleration (0.26σ), braking (0.10σ), and corner-

[4] In a related domain, Jackson and Schneider (2015) show how checklists improve worker productivity by serving as a "memory aid".

[5] Reported effect sizes are based on the within standard deviation of the sample used for estimation.

ing (0.20σ). Improvements tend to persist with a smaller magnitude in the ensuing weeks but fade out after about eight to nine weeks. This suggests that, as time progresses, coaching effects decay and drivers fall back into old driving habits.

There are important heterogeneous interaction effects between the peer-comparison treatments and on-the-road coaching. Comparing coached and uncoached drivers in the control group with general feedback, we find that coached drivers perform slightly better during the experimental period (0.12σ for fuel economy). Coached drivers in the treatment groups with targeted peer-comparison feedback, however, tend to perform worse in the experimental period compared to their coached colleagues in the control group (about 0.16σ for fuel economy). For uncoached drivers, in contrast, we find that drivers in the treatment groups improve on their driving behavior compared to their peers in the control group (0.12σ for fuel economy). These results show that it is important to consider interactions between interventions when designing corporate energy conservation schemes.

Our findings contribute to the broader societal challenge of combating unsustainable energy consumption practices. While there has been much progress in our understanding of non-financial incentives in residential energy consumption, more research is needed to explore the extent to which these insights generalize to firms (Gerarden, Newell and Stavins forthcoming, Gosnell et al. 2016). Our work aims to fill this gap by evaluating the immediate and delayed (interaction) effects of two tailored feedback programs on the driving behavior of bus drivers. It shows the potential of benefiting both the environment and the firm through simple behavioral adjustments by workers.

This paper proceeds as follows. Section 2 describes the field setting of the study. Section 3 elaborates on the research design and provides further details on the data and both feedback programs. Section 4 provides an empirical analysis of the feedback programs. Section 5 concludes.

2 Field Setting

2.1 Industry

Our field partner is Arriva, a European-wide passenger transport company operating various transport modes in the Dutch public transport industry.[6] This study focuses on buses, the firm's largest business unit. In the Netherlands, bus concessions are granted to companies by means of a tendering procedure (see Passenger Transport Act 2000). It provides a company with the exclusive rights to operate in a designated area for a number of years. As part of the procedure, the local government writes a statement of requirements with the envisioned goals and plans for the area. Environmental objectives feature prominently in these statements and stimulate interested firms to engage in green behavior.[7] The Dutch Ministry of Infrastructure noted that public transport is a "trend setter" in the area of sustainable technologies (MIVW 2010).

One such technology is an electronic on-board recorder (EOBR) which meticulously measures performance on several dimensions of driving behavior. For example, the version used by our partner records trip-level performance on fuel consumption and so-called 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. The EOBR technology enables precise monitoring and provides managers and researchers with a wealth of high-frequency data on worker productivity.

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. During each trip, the EOBR records "events" whenever an action by the driver is in excess of these thresholds. The number of events serves as a performance measure, with less 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.

[6] The Arriva Group is part of Deutsche Bahn. The company employs more than 60,000 people and annually delivers more than 2.2 billion passenger journeys in 14 European countries.

[7] In one recent statement, interested companies are requested to submit a sustainability plan in which they indicate how they make public transport in the concession area more environmentally friendly.

2.2 The Company

As part of a campaign on economical and comfortable driving behavior, Arriva Netherlands is currently in the process of installing new EOBRs in the entire fleet. The old technology reported large variations in fuel usage across buses, which the company attributes to differences in driving behavior. The EOBR data will therefore 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.

For this study, 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.[8] 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 currently against using financial incentives to reward good performances. In the past, drivers received no personal feedback and were not aware of being monitored.

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 concession. For five locations, virtually all routes are between cities and in rural areas. One location (the largest one) has a mixture of urban and rural routes. A special bus type, which runs on natural gas, operates on most of the urban trips. Routes are assigned based on timetables that do not vary much across time. Drivers are thus familiar with their routes and have had time to learn under different on-the-road circumstances. Timetables provide ample within-location variation in the type of trips, thereby ensuring that drivers face a more or less similar mixture of relatively easy and difficult trips.

[8] Based on the 2015 figure of check-ins with public transport cards.

3 Research Design

3.1 Research Setting

Figure 1: Timeline of the Study

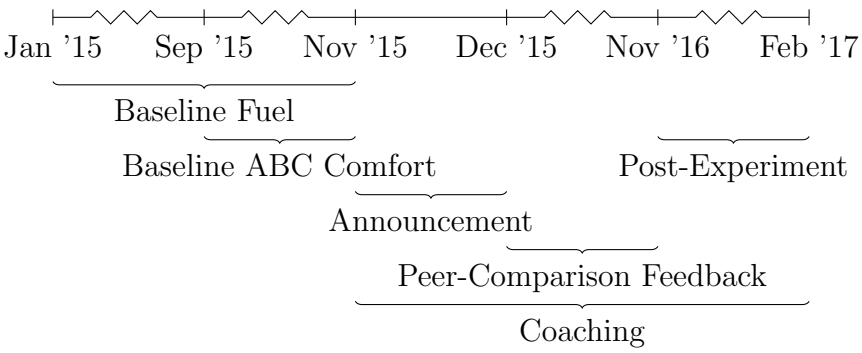


Figure 1 depicts the timeline of the study. First, we use the old on-board system to establish a long baseline of fuel consumption, starting in January 2015. At this stage, drivers are not informed about the upcoming feedback, nor that they are being monitored. The new EOBR system enables the collection of comfort dimensions baseline data in the months September and October 2015. Second, the period in between November 2015 and mid-December 2015 is used to disentangle the announcement effect from the feedback effect. The company held a kickoff event at the start of November 2015. At this event it was made clear to the drivers that they are being monitored and that they will receive feedback in the next month. In the third phase, from mid-December 2015 until November 2016, drivers receive their monthly feedback reports with peer-comparison feeedback. Finally, the post-experimental period involves a one-time notification to the drivers that the peer-comparison messages are no longer included in the reports.[9] The on-the-road coaching program starts around the kickoff event in November 2015.

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

[9]The 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. Hence, all drivers except those in the control condition. See Section 3.3 for further details on the experimental conditions.

these events is a convenient feature of our research setting. To rule out career concerns as an alternative explanation, drivers were informed during the announcement period that the feedback will not be used in formal evaluations. Apart from the feedback programs under consideration, no other incentives were used by the company to promote green behavior. 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.

3.2 Data Collection and Sample Construction

The EOBRs are installed in three bus types. Bus types 1 and 2 have diesel engines. Type 1 is most commonly used, accounting for 75% of trips performed by the average driver. Type 2 is mainly used for routes with a long travel distance and operates from selected base locations. The remaining trips are performed in bus type 3, which runs on natural gas, and is only part of the fleet in the urban area of the largest base location. Table 1 provides the start dates of data recording per bus type.

Table 1: Start Date of Data Recording per Bus Type

	Bus type 1	Bus type 2	Bus type 3
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

We perform a few steps to construct the final sample. The most important step is the removal of observations with no outcome data due to an absent bus identifier (290,737 obs; 22.73% of the full sample). We also drop observations with values that are clearly unreasonable, such as a fuel economy (kilometers per liter of fuel) of less than 1 or more than 8 (1,259 obs; 0.10%), a difference of more than 1 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%). The complete sample construction is detailed in the Appendix. The final sample consists of 842,845 trip-level observations, which we match with driver, trip, and daily weather characteristics.[10]

[10]Weather data are collected from a weather station located in the regional capital and is maintained by the Royal Netherlands Meteorological Institute (KNMI).

3.3 Treatment Variation: Feedback Reports

The company made clear during exploratory talks that it wants to target feedback messages at the poor-performing driving dimensions, arguing that the biggest gains in fuel-efficient and comfortable driving are to be made by focusing on the behaviors with the largest room for improvement. These messages are integrated in the monthly feedback reports. Our experiment varies the number and type of feedback messages and is best described as a natural field experiment (Harrison and List 2004) because drivers operate in a natural environment and are not informed about the experimental variation.

All 409 tenured drivers operating in the area 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 tenured drivers are randomly allocated to the experimental conditions, stratified along the dimensions of base location, gender, and years of service at the company.

In the experimental conditions, we construct reference groups in which driver performance on each comfort dimension is compared to colleagues with the same base location and treatment status.[11] This creates a natural and homogeneous comparison group for drivers in which competition is likely to generate strong incentives (Lazear and Rosen 1981, Delfgaauw et al. 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 per driver detailing his or her performance during the previous month. We use this information to determine how a driver performed compared to the reference group of peers and assign peer-comparison messages accordingly.[12] Negative (positive) messages are provided if a driver belongs to the bottom 50% (top 25%) of the reference group.

[11] This 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.

[12] The performance summary contains information about 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 each driving dimension is the weighted sum of the percentile ranks of the driver on each bus type. The weight is determined based on kilometers driven on a bus type as a share of total kilometers driven that month. All drivers within a reference group are ranked according to their final ranks in order to determine how a driver performed compared to peers.

In treatment 1 (1n/0p), 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 treatment 2 (1n/1p), 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 treatment 3 (3n/0p), drivers run the risk of receiving negative messages on all comfort dimensions. Using treatment 2 (1n/1p) 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 (with the messages integrated) is created around the 15th day of each feedback month and delivered 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.[13] 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. A sample feedback report is provided in Figure A1. The experimental conditions are summarized in Table 2.

Table 2: Experimental Conditions

Conditions	General feedback	Max # positive message(s)	Max # negative message(s)
Control	Yes	0	0
Treatment 1	Yes	0	1
Treatment 2	Yes	1	1
Treatment 3	Yes	0	3

[13]Working 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 3 provides a summary of the data per experimental condition and shows that the randomization is successful: both driver and trip characteristics are well balanced across the experimental groups. In particular, we observe no significant differences in driving behavior during the pre-experimental period. Drivers are on average 54 years old and work for 20 years at the company. Most drivers are male (89%). The average trip had a length of 31 km and was typically driven in rural areas (84%).

The tailored nature of the messages is illustrated in Table 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).[14] It highlights the flexible design of the treatments. Each treated driver is assigned an individualized message combination which points to behaviors that require attention. In treatment 1 (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 was in the top 25% or bottom 50% on a given driving dimension is shown in Figure A2. The figure plots 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 the driver received a feedback report. This gives an indication how often a driver was eligible for targeted messages. For many drivers it varied per round whether they were in the target groups. We also observe on each dimension 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.

This illustrates the motivation behind the messages. The detailed data allow for precise identification of good and bad performers in every feedback round. The messages are subsequently intended as a means to assist these drivers in offering guidance on where to improve (maintain) performance. They are updated every round to inform about progress and to avoid drivers from slacking off.

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

Table 3: Descriptive Statistics of Experimental Conditions

	Full sample		**C (0n/0p)**		**T1 (1n/0p)**		**T2 (1n/1p)**		**T3 (3n/0p)**	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Pre-experimental performance										
Fuel economy	4.06	0.31	4.05	0.30	4.09	0.32	4.05	0.33	4.05	0.31
Acceleration	12.30	6.43	12.38	6.65	11.91	6.59	12.28	6.27	12.61	6.27
Braking	3.00	4.39	3.20	4.63	2.98	4.60	2.59	3.66	3.25	4.63
Cornering	2.88	4.93	2.99	5.11	2.85	5.17	2.62	4.60	3.06	4.89
Demographics										
Year of birth	1962	8.28	1962	8.83	1962	8.62	1962	7.62	1962	8.12
Year of employment	1996	11.53	1996	11.65	1996	11.48	1996	11.32	1996	11.83
% share of FTE$\geq$0.9	76.28		75.73		74.51		79.41		75.49	
% share of female drivers	10.51		10.68		9.80		10.78		10.78	
Trip-specific variables										
Punctuality	-2.88	0.80	-2.80	0.86	-2.94	0.76	-2.80	0.86	-2.96	0.72
Distance traveled	31.36	13.21	31.32	13.99	31.89	12.59	31.14	12.84	31.10	13.53
Number of passengers	13.26	3.71	13.38	3.91	13.27	3.69	13.18	3.60	13.23	3.69
Number of bus stops	37.84	8.79	37.38	9.09	38.54	8.40	37.93	8.76	37.51	8.99
% share of rides:										
- Morning rush hours	19.45		19.34		20.53		18.64		19.31	
- Evening rush hours	19.53		19.99		18.20		20.55		19.39	
- Weekend	14.09		14.31		14.01		3.98		14.03	
- Holidays	9.99		9.88		10.27		9.71		10.10	
- Urban area	15.84		16.77		13.68		15.89		17.00	
- School	0.8		0.7		0.8		0.8		0.7	
% share of trips on bus types by the average driver										
Bus type 1	75.20		73.64		77.71		75.63		73.88	
Bus type 2	9.65		10.21		9.34		9.30		9.75	
Bus type 3	15.15		16.15		12.96		15.07		16.37	
Base locations (# drivers)										
Location 1	12		3		3		3		3	
Location 2	61		16		15		15		15	
Location 3	30		7		8		8		7	
Location 4	74		19		18		18		19	
Location 5	150		37		38		38		37	
Location 6	82		21		20		20		21	
Number of drivers	409		103		102		102		102	

Notes: data collected from electronic on-board recorders (EOBR) in buses and centralized databases with driver and trip characteristics. Fuel economy is defined as kilometers per liter of fuel. Performance on the comfort dimensions (acceleration, braking and cornering) is measured as the number of events per 10 kilometers. Less events means better driving behavior. 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 is based on check-ins with public transport cards. Bus types 1 and 2 have diesel engines and bus type 3 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. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

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

	T1 (1n/0p)		**T2 (1n/1p)**				**T3 (3n/0p)**			
Feedback round	0n/0p	1n/0p	0n/0p	1n/0p	0n/1p	1n/1p	0n/0p	1n/0p	2n/0p	3n/0p
	Panel A: incidence of messages									
December 2015	31%	69%	3%	53%	24%	20%	29%	26%	22%	23%
January 2016	27%	73%	7%	49%	22%	22%	29%	24%	17%	29%
February 2016	31%	69%	2%	49%	21%	28%	27%	27%	22%	24%
March 2016	30%	70%	3%	52%	23%	22%	32%	16%	24%	28%
April 2016	30%	70%	1%	53%	27%	19%	31%	18%	28%	23%
May 2016	29%	71%	2%	55%	27%	16%	28%	18%	34%	20%
June 2016	30%	70%	5%	51%	23%	22%	29%	19%	30%	22%
July 2016	26%	74%	3%	54%	24%	19%	25%	23%	26%	25%
August 2016	27%	73%	4%	47%	23%	25%	27%	24%	24%	24%
September 2016	32%	68%	2%	56%	26%	16%	28%	19%	34%	18%
October 2016	30%	70%	3%	52%	22%	23%	29%	22%	24%	25%
	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)	(100;100;100)
January 2016		(33;33;33)		(26;30;45)	(38;33;29)	(76;29;95)		(30;30;39)	(69;69;63)	(100;100;100)
February 2016		(20;41;39)		(34;43;23)	(55;15;30)	(67;59;74)		(44;20;36)	(52;81;67)	(100;100;100)
March 2016		(32;43;25)		(34;32;34)	(32;23;45)	(90;52;57)		(27;33;40)	(65;70;65)	(100;100;100)
April 2016		(33;38;30)		(27;37;35)	(42;23;35)	(56;72;72)		(35;24;41)	(63;74;63)	(100;100;100)
May 2016		(38;29;32)		(34;38;28)	(31;27;42)	(100;56;44)		(29;29;41)	(64;73;64)	(100;100;100)
June 2016		(40;35;25)		(31;37;33)	(45;27;27)	(67;62;71)		(28;28;44)	(68;71;61)	(100;100;100)
July 2016		(34;31;34)		(33;24;43)	(57;22;22)	(61;78;61)		(32;27;41)	(64;76;60)	(100;100;100)
August 2016		(38;30;33)		(29;36;36)	(32;32;36)	(88;58;54)		(43;30;26)	(61;61;78)	(100;100;100)
September 2016		(41;29;31)		(28;33;39)	(32;36;32)	(73;80;47)		(28;44;28)	(69;56;75)	(100;100;100)
October 2016		(47;25;27)		(39;27;35)	(38;24;38)	(50;86;64)		(50;25;25)	(55;68;77)	(100;100;100)

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. 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.

3.4 On-the-Road Coaching

Next to the decision to provide drivers with feedback reports, the company also initiated a coaching program. Six experienced drivers (spread over the base locations) were recruited as coaches based on their track record of driving behavior. All coaches participated in a training on how to approach drivers and to communicate feedback.

Since coaches are bus drivers themselves, there is only limited time available for coaching activities (about one day every two weeks).[15] Furthermore, the hop-on hop-off approach to on-the-road coaching makes prior and future sessions technically dependent on each other. For these reasons, it is not possible to randomly assign coaches to drivers. Coaches were requested at the beginning of the study to maintain a detailed log of their activities, allowing us to pinpoint when and how often drivers are coached.

We use the coach logs to determine for every coaching date the difference in mean baseline performance between drivers who received their first coaching versus uncoached colleagues who were also working from the same base location on that date. If the order of coaching is quasi-random and not based on pre-selected criteria, there should be no systematic difference between these two groups.[16] Table 5 provides evidence that the implementation of the coaching program exhibits a quasi-random order of phase-in. The table shows for multiple baseline performance measures the difference in means over all coaching dates. We observe for all locations combined that the difference is never significantly different from zero at the 5% level. Separating by base location, we obtain similar results for all but one base location. Note, however, that the differences at this location are similar to what is found for other locations.

Table A2 repeats the analysis for non-performance-related descriptives. Importantly, there are no significant differences in the share of experimental groups. This is convenient as it shows that coaches did not select according to who receives peer-comparison messages. Balance is also observed on other driver and trip characteristics. At two locations drivers with a greater share of trips during morning (evening) rush hours have a slightly greater (smaller) chance of being coached. This is likely to be planning-related: most

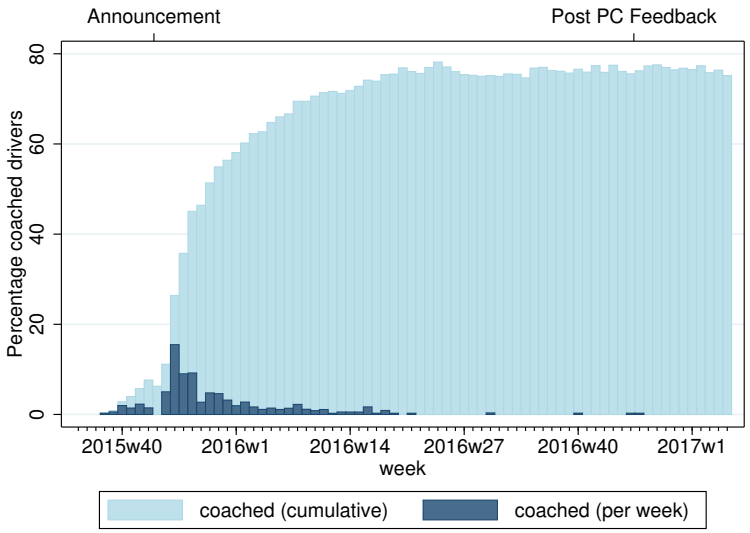
[15] Coaches can decide which day they use for coaching. They vary the day of the week such that every driver has a chance of being coached.

[16] Duflo, Glennerster and Kremer (2007) argue that a randomized order of phase-in is often the fairest way of implementing a program over time in contexts where it is not acceptable for individuals to receive no support.

coaches start early and then quit at some point.

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 passengers 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. Drivers are treated as equals and feedback is delivered in a constructive and positive manner.

Figure 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 around 1 November 2015 and first distributed as reports after 15 December 2015, with follow-up reports in each subsequent month. Peer-comparison messages were removed from the reports from November 2016 on.

Figure 2 shows that most drivers received their first coaching in the weeks following the kickoff event in November 2015. 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

Table 5: Quasi-Random Phase-In of Coaching - Baseline Performance of Coached and Uncoached Drivers

	Fuel economy			**Acceleration**			**Braking**			**Cornering**		
Base location	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC
Location 1	4.03	4.12	-0.09	11.26	10.89	0.37	1.95	1.77	0.18	0.85	0.50	0.35
Location 2	4.16	4.21	-0.04	11.54	11.07	0.47	1.87	1.79	0.08	1.46	1.39	0.07
Location 3	4.18	4.11	0.07	11.25	11.99	-0.73	1.53	1.74	-0.21	1.10	1.42	-0.32
Location 4	4.21	4.21	-0.00	10.53	10.85	-0.32	1.58	1.69	-0.11	0.83	1.15	-0.33
Location 5	3.84	3.79	0.05	18.43	18.41	0.02	8.08	8.18	-0.10	1.22	1.08	0.14
Location 6	4.27	4.21	0.06**	9.68	10.37	-0.69***	1.57	1.64	-0.07	1.04	1.31	-0.27**
All locations	4.08	4.05	0.03*	13.16	13.38	-0.22	3.72	3.80	-0.08	1.12	1.22	-0.10

Notes: coach logs are used to determine for every coaching date the difference in mean baseline performance between drivers who received their first coaching (C) versus uncoached colleagues (NC) who were also working from the same base location on that date. This table reports the mean over all coaching dates per base location. Fuel economy is kilometers per liter of fuel. Performance on the ABC comfort dimensions is measured as the number of events per 10 kilometers. Less events means better driving behavior. Stars indicate that the mean difference is significantly different from zero. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

feedback report in mid-December 2015. In a select number of cases, drivers participated in additional coaching sessions (55 drivers, 18% of all coached drivers). We control for these extra coaching sessions in our analyses.

The figure also highlights two other issues related to the coaching program. 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. In the regression analysis, we change for these drivers the post-announcement date from 1 November 2015 to their respective coaching dates.

4 Results

4.1 Feedback reports

Drivers are considered to be treated when a feedback report is created and delivered via the team manager or pigeonhole. A post-feedback indicator takes on the value 1 after receiving the first feedback report, 0 otherwise. This definition makes no selection on the actual receipt of the peer-comparison messages. From a policy perspective this is useful because it captures the aggregate performance of the treatments when applied to an eligible population (see also Allcott 2011). For each driver we can determine when a new report is created and delivered. The start of the post-feedback period may differ per driver due to a possible absence in the prior month to the one 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.

In order to identify the aggregate performance of the treatments, we make use of a difference-in-differences regression specification that models our measures of driving behavior Y_{it} conditional on treatment status T_i, an indicator for the post-feedback period $postfeedback_{it}$, a vector of control variables X_{it}, driver fixed effects μ_i, day fixed effects υ_d, and bus type fixed effects κ_b:

$$Y_{it} = T_i \cdot postfeedback_{it} \cdot \tau + X_{it} \cdot \theta + \mu_i + \upsilon_d + \kappa_b + \upsilon_d \cdot \kappa_b + \epsilon_{it} \tag{1}$$

Day fixed effects are included rather than a common post-feedback indicator in order to account for driver-specific starting dates of the post-feedback period. By 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. As control variables, we include a no-report indicator, travel distance, punctuality (difference between actual and planned driving time), and the number of passengers and bus stops. We also incorporate dummies for morning (7:00-10:00) and evening (16:00-19:00) rush hours, trips in urban areas, fill-in rides, and trips that were driven in a smaller or larger bus. We control for daily weather conditions (average temperature/wind and total rainfall) in robustness check specifications without day fixed effects. The model is estimated with robust standard errors, clustered by drivers so as to account for within-driver correlation patterns in the error term ϵ_{it} (Bertrand, Duflo and Mullainathan 2004).

The results are presented in Table 6. For ease of exposition, we restrict attention to fuel economy and acceleration and refer to the Appendix for other driving dimensions. Results are shown for all trajectories as well as for urban and rural areas separately. In the post-feedback period, all urban trips were performed in bus type 3 and there are therefore no post-feedback outcomes for fuel economy on urban trajectories. We prefer to think of urban trips as a separate sub-sample. As discussed in Section 2.2, almost all urban trips were in the vicinity of base location 5 (about 98%) and driven in bus type 3 (about 96%), which runs on natural gas, has a distinct set of thresholds for the ABC dimensions, and operates exclusively from location 5. To facilitate comparisons across driving behaviors, we will thus focus mainly on rural trips in our discussion of the results.

In general, the table shows that the peer-comparison messages do not further improve driving behavior after controlling for the "general feedback effect" of the control group. The point estimates across treatments for an aggregate measure of driving behavior (fuel economy) are small in size, about 0.007 (0.015σ)[17], and are (jointly) not statistically significant.[18] Similar qualitative results are obtained for acceleration as a disaggregated driving behavior. Table A3 in the Appendix indicates that braking and cornering are also not affected by the peer-comparison messages at a conventional 5% significance level.

[17]Reported effect sizes are based on the within standard deviation of the sample used for estimation.

[18]In a Wald test, we cannot reject the null hypothesis that the treatment effects are jointly equal to zero ($F = 0.20$, $p = 0.896$).

Table 6: Targeted Peer-Comparison Feedback Effects on Driving Performance

	Fuel economy					**Acceleration**				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Post-announcement	0.095***	0.095***				-1.674***	-1.674***			
	(0.006)	(0.006)				(0.181)	(0.181)			
Post-feedback	-0.018*	-0.017*				-0.508**	-0.508**			
	(0.010)	(0.010)				(0.225)	(0.225)			
T1 (1n/0p)	0.013					-0.143				
	(0.023)					(0.287)				
T2 (1n/1p)	0.008					0.288				
	(0.027)					(0.348)				
T3 (3n/0p)	0.006					0.288				
	(0.024)					(0.301)				
Post-feedback x T1 (1n/0p)	0.013	0.013	0.011	n.a.	0.010	0.236	0.236	0.255	1.460	-0.084
	(0.016)	(0.016)	(0.015)		(0.015)	(0.303)	(0.303)	(0.306)	(0.948)	(0.215)
Post-feedback x T2 (1n/1p)	0.002	0.002	0.003	n.a.	0.004	-0.398	-0.402	-0.413	-1.577	-0.009
	(0.013)	(0.013)	(0.013)		(0.013)	(0.477)	(0.477)	(0.459)	(1.702)	(0.205)
Post-feedback x T3 (3n/0p)	0.008	0.008	0.007	n.a.	0.008	-0.245	-0.247	-0.278	-0.212	-0.239
	(0.013)	(0.013)	(0.013)		(0.013)	(0.418)	(0.418)	(0.395)	(1.256)	(0.209)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Day fixed effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Trajectories	All	All	All	Urban	Rural	All	All	All	Urban	Rural
Number of drivers	405	405	405	234	394	398	398	398	120	389
Number of trip-level observations	498,667	498,667	498,667	7,683	490,984	473,042	473,042	473,042	135,337	337,705

Notes: identification of the treatment effects on driving performance. Standard errors in parentheses. The time period under consideration is from 1 January 2015 until 15 November 2016, when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. 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. Post-announcement is a dummy variable with value 1 in the period after 1 November 2015 (feedback announcement), 0 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 after 15 December 2015. The dependent variables are fuel economy (km/liters) and for acceleration it is the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, trips that were driven in a smaller or larger bus, and daily weather conditions (average temperature/wind and total rainfall). A no-report indicator is also included and captures drivers operating after 15 December 2015 (first feedback round) but who have not yet received their first report. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

From the negative (positive) sign of the no-report indicator (not shown) on fuel economy (acceleration), we conjecture that drivers in general do respond to the feedback reports but that there is no additional overall effect of the peer-comparison messages. Fuel economy among no-report drivers is 0.06 lower compared to post-feedback drivers in the control group, which constitutes a small effect of about 0.13σ.

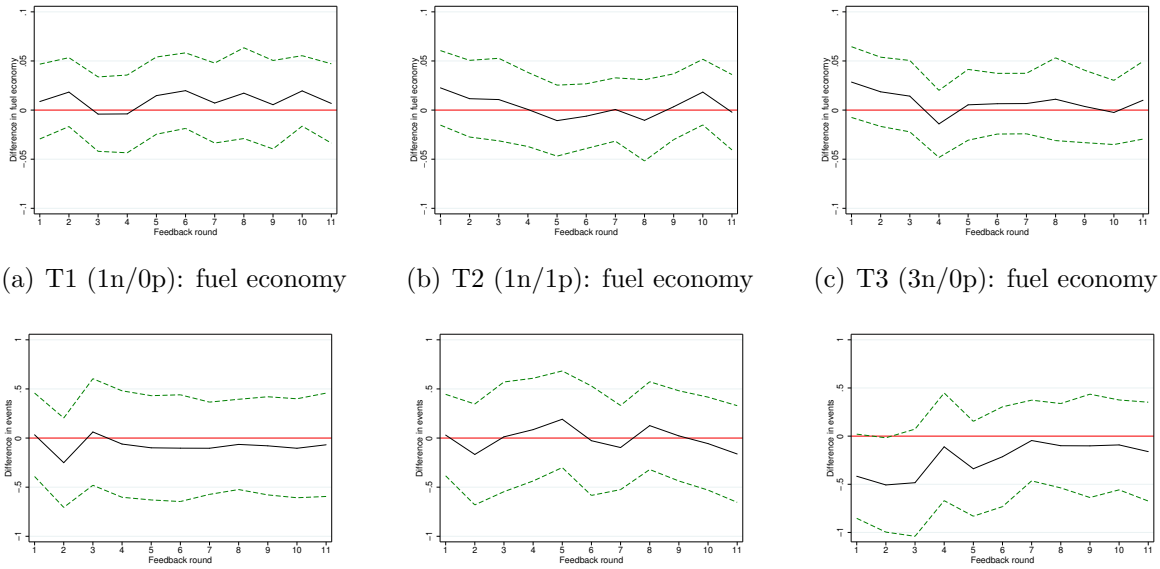
The absence of an effect of peer-comparison feedback on conservation efforts among workers is consistent with findings from another recent study conducted in a similar setting (Blader et al. 2016). In that study, however, it is noted that absent aggregate effects may mask temporal effects and improved performance among sub-groups of drivers. By identifying feedback rounds and post-coaching rides, we can examine how peer-comparison messages change driving over time and whether effects differ for (un)coached drivers.

Figure 3 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 contained a text message notifying all treated drivers that they will no longer receive peer-comparison messages. We observe that drivers in treatment 3 (3n/0p) perform slightly better on acceleration in the first few feedback rounds, about half an event less per 10 kilometers (0.12σ) but this effect fades out over time. The general pattern, however, is that there are no temporal effects of the peer-comparison messages on driving behavior.

Table A4 compares pre-experimental to post-experimental driving performance. We observe that treated and control drivers perform very similar in the post-experimental period (15 November 2016 - 31 January 2017). This indicates that treated drivers did not habituate to a more efficient and comfortable driving style on top of what is already observed among control drivers. It also suggests that treated drivers did not respond to the notification of a withdrawal of peer-comparison messages in the feedback reports.[19]
Thus far we have disregarded any differences between coached and uncoached drivers. Given the more or less simultaneous introduction of coaching sessions and feedback reports, however, it is likely that drivers compare the feedback from both programs and may adjust their effort accordingly. First, we aim to quantify the independent effect of a single coaching session and then proceed with a discussion of possible interaction effects.

[19]Results on temporal and post-experimental effects are similar for braking and cornering. In the interest of space, we do not include these results in the Appendix but they are available upon request.

Figure 3: Temporal Effects Targeted Peer-Comparison Feedback



(a) T1 (1n/0p): fuel economy (b) T2 (1n/1p): fuel economy (c) T3 (3n/0p): fuel economy

(d) T1 (1n/0p): acceleration (e) T2 (1n/1p): acceleration (f) T3 (3n/0p): acceleration

Notes: treatment effects per feedback round based on rural trips. The dotted green lines indicate 95% confidence intervals. The time period is from 1 January 2015 until 15 November 2016, when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. Fuel economy is defined as kilometers per liter of fuel. Acceleration is measured as the number of events per 10 kilometers (less events means better driver behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for driver, day, bus type, and bus type x day fixed effects as well as for punctuality (difference between actual and planned driving time), morning and evening rush hours, smaller and larger buses, fill-in trips, trip distance, number of passengers and bus stops, and drivers operating after 15 December 2015 (first feedback round for most drivers) but who have not yet received their first report.

4.2 On-The-Road Coaching

In order to identify the effect of a single on-the-road coaching session, we restrict our attention to the first coaching moment of each driver and control for the small number of additional coaching sessions. We observe the driving behavior of coached and uncoached drivers (coaches are excluded from this sample) before and after the session and therefore make use of a difference-in-differences approach to identify the coaching effect. Our variable of interest, $postcoaching_{it}$, identifies the day of first coaching and the period thereafter for each driver. Day fixed effects υ_d are included to absorb general day-to-day variation in driving behavior that is not due to coaching or explained by the control variables. We use the same set of control variables as in the previous section. The corre-

sponding regression is estimated with robust standard errors, clustered by drivers, and is specified as follows:

$$Y_{it} = postcoaching_{it} \cdot \beta + X_{it} \cdot \theta + \mu_i + \upsilon_d + \kappa_b + \upsilon_d \cdot \kappa_b + \epsilon_{it} \tag{2}$$

The results are provided in Table 7. The first three columns report per driving dimension the overall effects on different trajectories. Focusing on rural trips, we observe that a single coaching session (marginally) improves fuel economy (0.04σ, $p = 0.057$) and acceleration (0.11σ, $p < 0.001$). The last three columns show that the gains from coaching are mainly achieved in the first weeks after coaching. In these columns, we replace the post-coaching dummy with indicators for the day of coaching and the first weeks thereafter. The largest improvements are observed during the coaching day: fuel economy goes up (0.23σ, $p < 0.001$) and acceleration events go down (0.26σ, $p < 0.001$). Improvements persist in the ensuing weeks, albeit with a smaller magnitude. Table A5 provides the results for braking and cornering. Braking events only go down during the day of coaching (0.10σ, $p = 0.002$), with little evidence of persistence in the ensuing weeks. Cornering also improves during the coaching day (0.20σ, $p < 0.001$) and there is some evidence of short-run persistence in the weeks following the coaching session.

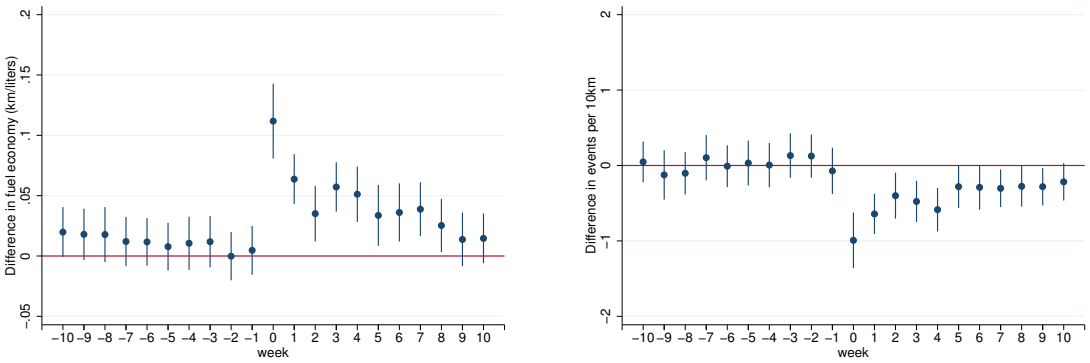
Figure 4 further elaborates on temporal effects by plotting coaching effects on rural trips in the weeks that are in close proximity to the coaching date. We observe strong improvements during and in the wake of coaching. These gains in fuel economy and acceleration persist for about eight to nine weeks. Improvements are less precisely estimated at the end of the 10-week interval. This suggests that, as time progresses, coaching effects decay and drivers seem to fall back into old driving habits. We observe no differences in driving behavior in the ten weeks prior to coaching, which lends support to our earlier conclusion that the selection for a coaching session is random and not based on prior performance.

Table 7: On-The-Road Coaching Effects on Driving Performance

	Fuel economy						**Acceleration**					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Post-coaching	0.019*	n.a.	0.019*				-0.547***	-0.911	-0.438***			
	(0.010)		(0.010)				(0.176)	(0.755)	(0.122)			
Day of first coaching				0.103***		0.103***				-1.657***	-3.274**	-1.046***
				(0.016)		(0.016)				(0.427)	(1.295)	(0.173)
Days after first coaching:												
1-7				0.055***	n.a.	0.055***				-0.639**	-0.832	-0.701***
				(0.010)		(0.010)				(0.251)	(0.914)	(0.122)
8-14				0.027**	n.a.	0.026**				0.105	1.254	-0.465***
				(0.012)		(0.012)				(0.295)	(1.041)	(0.141)
15-21				0.048***	n.a.	0.048***				-0.164	0.350	-0.543***
				(0.011)		(0.011)				(0.248)	(0.941)	(0.134)
22-28				0.042***	n.a.	0.042***				-0.327	0.034	-0.656***
				(0.013)		(0.013)				(0.265)	(0.843)	(0.149)
> 28				0.003	n.a.	0.004				-0.685***	-2.045**	-0.266*
				(0.012)		(0.012)				(0.200)	(0.799)	(0.149)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Trajectories	All	Urban	Rural	All	Urban	Rural	All	Urban	Rural	All	Urban	Rural
Number of drivers	399	230	383	399	230	383	391	112	376	391	112	376
Number of trip-level observations	352,253	7,533	344,720	352,253	7,533	344,720	262,332	75,205	187,127	262,332	75,205	187,127

Notes: identification of on-the-road coaching effects on driving performance. Standard errors in parentheses. The time period under consideration is the period for which we have complete logs available from all coaches (1 January 2015 - 30 April 2016). Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. The dependent variables are fuel economy (kilometers per liter of fuel) and for acceleration it is the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for post-coaching rides after an additional coaching session, travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, and trips that were driven in a smaller or larger bus. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure 4: Temporal Effects On-The-Road Coaching



(a) Fuel economy

(b) Acceleration

Notes: driving performance in the weeks before and after coaching based on rural trips. The vertical spikes indicate 95% confidence intervals. The dependent variables are fuel economy (km/liters) and for acceleration it is the number of events per 10 kilometers (less events means better driving behavior). Each graph plots the regression coefficients of dummy variables indicating the day of coaching (normalized to zero) and the weeks that are in close proximity to this event. Regressions are estimated with robust standard errors (clustered by drivers) and control for post-coaching rides after an additional coaching session, travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, fill-in rides, and trips that were driven in a smaller or larger bus. Coaches are excluded from the sample.

4.3 Interaction Effects

Table 8 documents the interaction effects between the two feedback programs. Comparing coached and uncoached drivers in the control group with general feedback, we find that coached drivers perform slightly better during the experimental period. Fuel economy is 0.053 higher (0.12σ, $p = 0.001$) and excessive acceleration decreased with about half an event less per 10 kilometers (0.13σ, $p = 0.010$).[20]

Importantly, however, we observe that coached drivers in treatments 1 (1n/0p) and 3 (3n/0p) perform significantly worse in the experimental period compared to the coached drivers in the control group, with a lower fuel economy of about 0.077 (0.17σ, $p = 0.008$) and 0.065 (0.14σ, $p = 0.020$), respectively. The adverse direction of the effects is similar for the ABC dimensions but these effects are, in general, less precisely estimated. The negative interaction effects indicate that peer-comparison messages have an adverse effect

[20]Table A6 shows that there is also improved braking (0.11σ, $p = 0.018$). There are no significant differences in cornering.

on performance when recipients already received tailored coaching in the past.

For uncoached drivers, in turn, we find that drivers in treatments 1 (1n/0p) and 3 (3n/0p) tend to improve on their driving behavior compared to drivers in the control group. Fuel economy is about 0.052 higher (0.12σ). Improvements are also observed on the ABC dimensions but these, again, are less precisely estimated. Our results stress the importance of taking into account interaction effects when designing and implementing multiple energy conservation interventions.

5 Conclusion

Increasing conservation efforts among workers is seen as low-hanging fruit in the battle to reduce corporate energy consumption. In light of institutional constraints that hinder the use of pay-for-performance plans, however, many firms are confronted with the question how to motivate workers to save energy when a financial incentive to do so is absent. This has triggered much recent interest in the design and evaluation of non-financial incentives in corporate conservation schemes.

In the transport industry, recent innovations in on-board monitoring technology open up novel opportunities for tailoring and evaluating non-financial incentives, allowing firms to identify high users and to tailor feedback to the underlying sources of energy consumption. In this paper, we collaborate with a large public transport company and benefit from their recent installation of on-board computers in the entire bus fleet. In a natural field experiment, we evaluate targeted peer-comparison feedback by randomly assigning individualized reports with varying numbers of peer-comparison messages to more than 400 bus drivers. Furthermore, we collect coach logs to quasi-experimentally evaluate the company's initiative to coach drivers during on-the-road sessions.

Our analysis reveals that there are important interaction effects between the two feedback programs. While the aggregate performance of the peer-comparison messages is not significantly different from a control group with general feedback, we observe that coached and uncoached drivers respond differently to the messages. Uncoached drivers show some improvement after being exposed to peer comparisons but this appears to be offset by their coached colleagues, who perform worse in the experimental period compared to coached drivers in the control condition. On-the-road coaching itself positively affects

multiple driving behaviors but the effects diminish over time.

These findings contribute to a growing body of literature on non-financial interventions aimed at energy conservation. Most of these studies focus on households and relatively little attention has been paid to firms. Recent studies, therefore, evaluate these interventions in corporate settings and provide us with new insights about worker motivation in relation to conservation efforts. This paper complements these studies and suggests that there is great potential in reducing corporate energy consumption by designing and implementing policies that target workers rather than the firm as a whole.

Our research points to several directions for future research. First, using monitoring technology to trace performance differentials to its underlying sources creates ample opportunity for tailoring incentives. In this paper, we use data on various direct measures of driving behavior to inform drivers where they perform well or poor compared to peers. Firms increasingly have these types of data available and we reckon that designing and evaluating other data-driven incentives could yield fruitful research. Second, we document important interaction effects between peer-comparison feedback and on-the-road coaching and believe that more research can be done that investigates interactions between non-financial incentives. Finally, and more in general, how conservation efforts can be stimulated when someone else pays the bill is a question in need of more answers.

References

Abrahamse, Wokje, Linda Steg, Charles Vlek, and Talib Rothengatter, "A Review of Intervention Studies Aimed at Household Energy Conservation," *Journal of Environmental Psychology*, 2005, *25* (3), 273–291.

Allcott, Hunt, "Social Norms and Energy Conservation," *Journal of Public Economics*, 2011, *95* (9-10), 1082–1095.

____ **and Michael Greenstone**, "Is There an Energy Efficiency Gap," *Journal of Economic Perspectives*, 2012, *26* (1), 3–28.

____ **and Sendhil Mullainathan**, "Behavior and Energy Policy," *Science*, 2010, *327* (5970), 1204–1205.

____ **and Todd Rogers**, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, 2014, *104* (10), 3003–3037.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee, "Awards Unbundled: Evidence from a Natural Field Experiment," *Journal of Economic Behavior & Organization*, April 2014, *100*, 44–63.

Azmat, Ghazala and Nagore Iriberri, "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students," *Journal of Public Economics*, 2010, *94* (7-8), 435–452.

Baker, George P. and Thomas N. Hubbard, "Make Versus Buy in Trucking: Asset Ownership, Job Design, and Information," *American Economic Review*, 2003, *93* (3), 551–572.

Barankay, Iwan, "Rank Incentives: Evidence from a Randomized Workplace Experiment," *Working paper*, 2012.

Benabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates," *The Quarterly Journal of Economics*, 2004, *119* (1), 249–275.

Blader, Steven, Claudine Gartenberg, and Andrea Prat, "The Contingent Effect of Management Practices," *Working paper*, 2016.

Blanes i Vidal, Jordi and Mareike Nossol, "Tournaments without Prizes: Evidence from Personnel Records," *Management Science*, 2011, *57* (10), 1721–1736.

Brynjolfsson, Erik. and Kristina McElheran, "The Rapid Adoption of Data-Driven Decision-Making," *American Economic Review*, 2016, *106* (5), 133–139.

Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke, "Tournament Incentives in The Field: Gender Differences in The Workplace," *Journal of Labor Economics*, 2013, *32* (2), 305–326.

Dohmen, Thomas, Armin Falk, Klaus Fliessbach, Uwe Sunde, and Bernd Weber, "Relative versus absolute income, joy of winning, and gender: Brain imaging evidence," *Journal of Public Economics*, 2011, *95* (279-285).

Duflo, Esther, Rachel Glennerster, and Michael Kremer, *Using Randomization in Development Economics Research: A Toolkit*, Vol. 4, North-Holland,

EIA, U.S., "Annual Energy Outlook 2015," Technical Report, U.S. Energy Information Administration 2015.

Eriksson, Tor, Anders Poulsen, and Marie Claire Villeval, "Feedback and Incentives: Experimental Evidence," *Labour Economics*, 2009, *16* (6), 679–688.

Ferraro, Paul J. and Michael K. Price, "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment," *The Review of Economics and Statistics*, 2013, *95* (1), 64–73.

Freeman, Richard B., "Union Wage Practices and Wage Dispersion within Establishments," *Industrial and Labor Relations Review*, 1981, *36* (1), 3–21.

Gerarden, Todd D., Richard G. Newell, and Robert N. Stavins, "Assessing the Energy-Efficiency Gap," *Journal of Economic Literature*, forthcoming.

Golman, Russell, David Hagmann, and George Loewenstein, "Information Avoidance," *Journal of Economic Literature*, 2017, *55* (1), 96–135.

Gosnell, Greer K., John. A List, and Robert Metcalfe, "A New Approach to an Age-Old Problem: Solving Externalities By Incenting Workers Directly," *NBER Working Paper Series*, 2016, (WP nr. 22316).

Hahn, Robert., Robert D. Metcalfe, David Novgorodsky, and Michael K. Price, "The Behavioralist as Policy Designer: The Need to Test Multiple Treatments to Meet Multiple Targets," *NBER Working Paper Series.*, 2016, (No. 22886).

Harrison, Glenn W. and John A. List, "Field Experiments," *Journal of Economic Literature*, 2004, *42* (4), 1009–1055.

Holladay, Scott J., Jacob LaRiviere, David M. Novgorodsky, and Michael K. Price, "Asymmetric Effects of Non-pecuniary Signals on Search and Purchase Behavior for Energy-Efficient Durable Goods," *NBER Working Paper Series*, 2016, (No. 22939).

ICC, "From Laboratory to Road: A Comparison of Official and 'Real-World' Fuel Consumption CO2 Values for Cars in Europe and United States," Technical Report, International Council on Clean Transportation 2013.

IEA, "Improving the Fuel Economy of Road Vehicles: A Policy Package," Technical Report, International Energy Agency 2012.

Jackson, C. Kirabo and Henry S. Schneider, "Checklists and Worker Behavior: A Field Experiment," *American Economic Journal: Applied Economics*, 2015, *7* (4), 136–168.

Kuhnen, Camelia M. and Agnieszka Tymula, "Feedback, Self-Esteem, and Performance in Organizations," *Management Science*, 2012, *58* (1), 94–113.

Lam, Chak Fu, D. Scott DeRue, Elizabeth P. Karam, and John R. Hollenbeck, "The impact of feedback frequency on learning and task performance: challenging the "more is better" assumption," *Organizational Behavior and Human Decision Processes*, 2011, *116* (2), 217–228.

Lazear, Edward P. and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 1981, *89* (5), 841–864.

MIVW, "Public Transport in the Netherlands," Technical Report, Ministry of Transport, Public Works and Water Management 2010.

Moldovanu, Benny, Aner Sela, and Xianwen Shi, "Contests for Status," *Journal of Political Economy*, 2007, *115* (2), 338–363.

Ryan, Richard M. and Edward L. Deci, "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist*, 2000, *55* (1), 68–78.

Tran, Anh and Richard Zeckhauser, "Rank as an Inherent Incentive: Evidence from a Field Experiment," *Journal of Public Economics*, 2012, *96* (9-10), 645–650.

A Sample Construction

Table A1: Cleaning Steps for Sample Construction

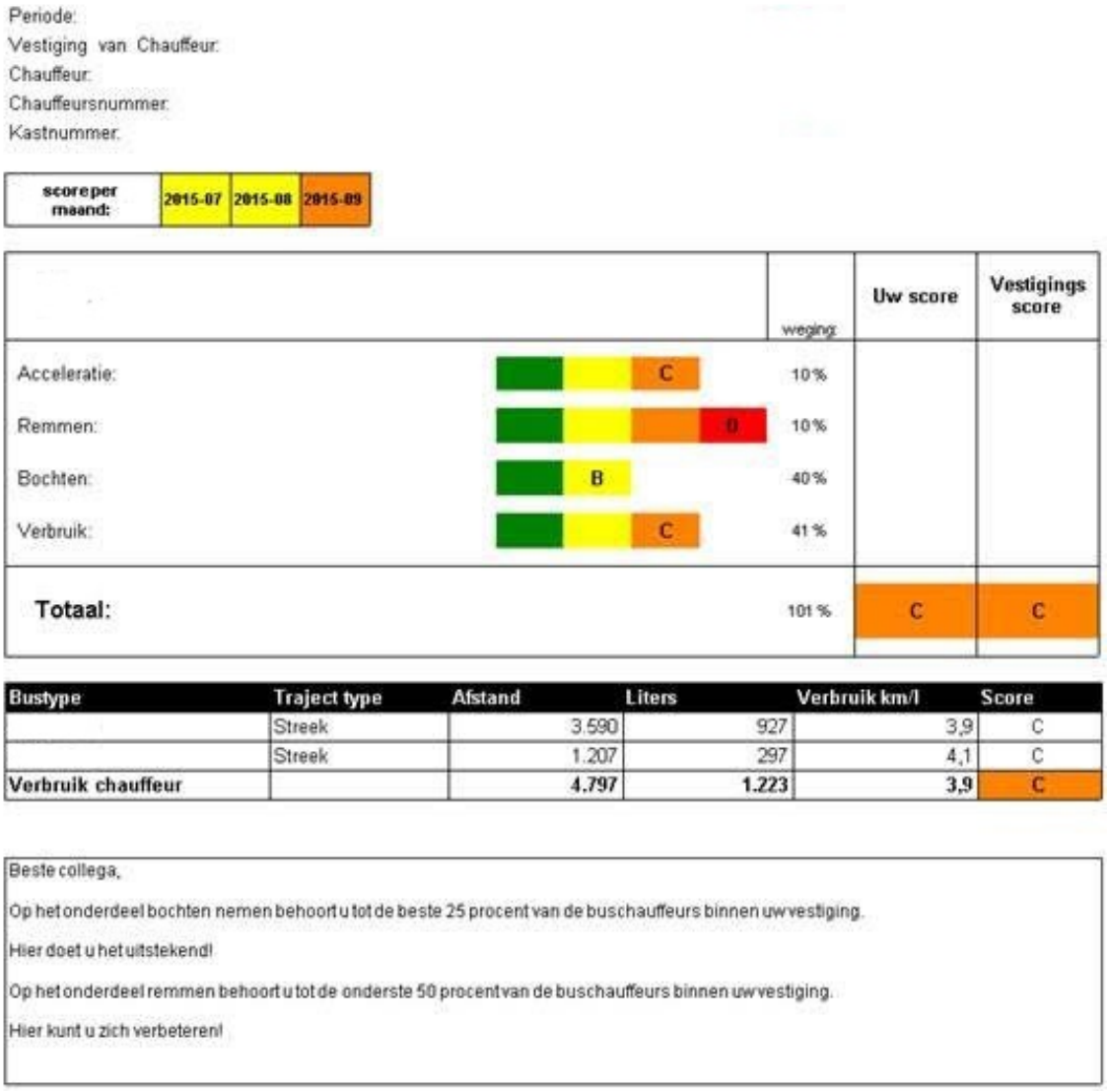
		% share of full sample
Full sample	1,278,913	
Reason for dropping observation:		
Duplicate observation (in terms of all variables)	(6,762)	0.5%
No bus identifier	(290,737)	22.73%
Bus type not eligible for EOBR	(34,870)	2.73%
Error message from EOBR	(31,118)	2.46%
Within-driver obs. with the same departure 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%
Unreasonable outcomes of dependent variables:		
- Fuel economy: less than 1 or more than 8	(1,259)	0.10%
- ABC dimensions: more than 5 SDs above the mean	(4,003)	0.31%
Final sample	842,845	

Notes: fuel economy is measured as kilometers per liter of fuel. The ABC comfort dimensions are the number of events per 10 kilometers (less events means better driving behavior). Punctuality is the difference between actual and planned driving time.

B Sample Feedback Report

Confidential information (related to the driver and company) has been removed from the feedback report.

Figure A1: Sample Feedback Report



Periode:
Vestiging van Chauffeur:
Chauffeur:
Chauffeursnummer:
Kastnummer:

score per maand:	2015-07	2015-08	2015-09

	weging	Uw score	Vestigings score
Acceleratie: C	10 %		
Remmen: D	10 %		
Bochten: B	40 %		
Verbruik: C	41 %		
Totaal:	101 %	C	C

Bustype	Traject type	Afstand	Liters	Verbruik km/l	Score
	Streek	3.590	927	3,9	C
	Streek	1.207	297	4,1	C
Verbruik chauffeur		**4.797**	**1.223**	**3,9**	**C**

Beste collega,

Op het onderdeel bochten nemen behoort u tot de beste 25 procent van de buschauffeurs binnen uw vestiging.

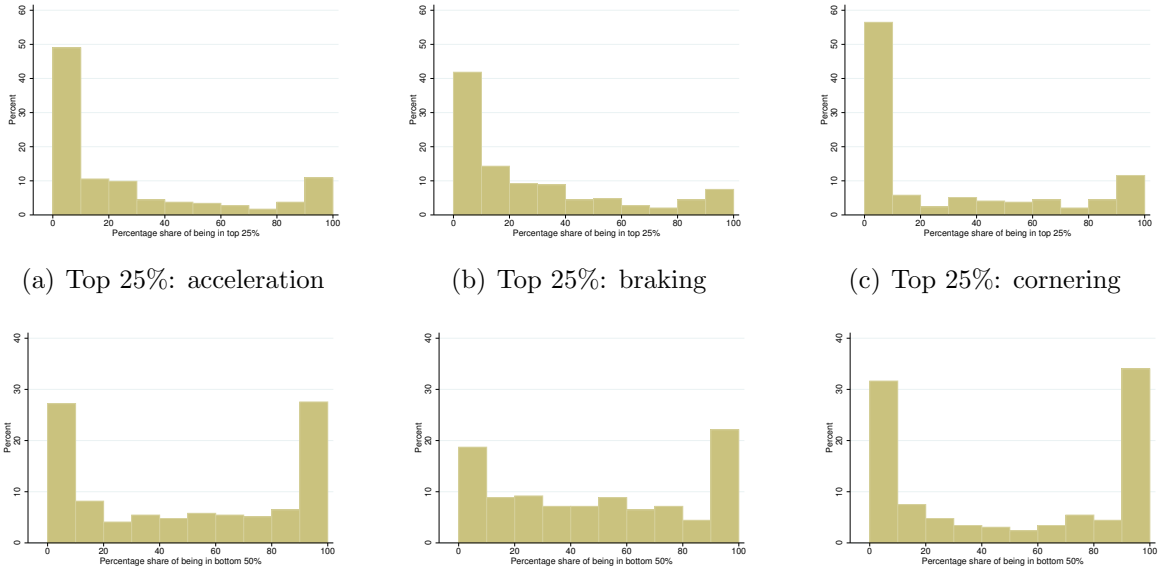
Hier doet u het uitstekend!

Op het onderdeel remmen behoort u tot de onderste 50 procent van de buschauffeurs binnen uw vestiging.

Hier kunt u zich verbeteren!

C Eligibility for Targeted Peer-Comparison Messages

Figure A2: Share of Feedback Rounds in Top 25% or Bottom 50% for Treated Drivers



(a) Top 25%: acceleration (b) Top 25%: braking (c) Top 25%: cornering

(d) Bottom 50%: acceleration (e) Bottom 50%: braking (f) Bottom 50%: cornering

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.

Table 8: Interaction Effects Targeted Peer-Comparison Feedback and On-The-Road Coaching

	Fuel economy			**Acceleration**		
	(1)	(2)	(3)	(1)	(2)	(3)
Post-feedback × T1 (1n/0p)	0.053**	n.a.	0.052**	-0.442	-1.171	-0.419
	(0.024)		(0.024)	(0.331)	(1.396)	(0.256)
Post-feedback × T2 (1n/1p)	0.032	n.a.	0.034	-0.313	1.498	-0.180
	(0.027)		(0.027)	(0.364)	(1.538)	(0.322)
Post-feedback × T3 (3n/0p)	0.053**	n.a.	0.052**	-0.728**	-0.298	-0.759**
	(0.025)		(0.025)	(0.356)	(1.182)	(0.294)
Post-coaching × Post-feedback	0.053***	n.a.	0.053***	-0.782**	-2.793**	-0.514**
	(0.016)		(0.016)	(0.322)	(1.198)	(0.198)
Post-coaching × Post-feedback x T1 (1n/0p)	-0.077***	n.a.	-0.077***	1.162**	3.313*	0.548*
	(0.029)		(0.029)	(0.483)	(1.697)	(0.329)
Post-coaching × Post-feedback × T2 (1n/1p)	-0.045	n.a.	-0.047	0.097	-2.559	0.408
	(0.031)		(0.031)	(0.633)	(1.576)	(0.355)
Post-coaching × Post-feedback × T3 (3n/0p)	-0.065**	n.a.	-0.065**	0.608	0.576	0.561
	(0.028)		(0.028)	(0.558)	(1.779)	(0.351)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Trajectories	All	Urban	Rural	All	Urban	Rural
Number of drivers	352,253	7,533	344,720	262,332	75,205	187,127
Number of trip-level observations	399	230	383	391	112	376

Notes: identification of the interaction effects between the two feedback programs. Standard errors in parentheses. The time period under consideration is the period for which we have complete coach logs available (1 January 2015 - 30 April 2016). Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. 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 after 15 December 2015. The dependent variables are fuel economy (km/liters) and for acceleration it is the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, and trips that were driven in a smaller or larger bus. A no-report indicator is also included and captures drivers operating after 15 December 2015 (first feedback round) but who have not yet received their first report. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A2: Quasi-Random Phase-In of Coaching - Non-Performance Descriptives of Coached and Uncoached Drivers

	All locations			**Location 1**			**Location 2**			**Location 3**			**Location 4**			**Location 5**			**Location 6**		
	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC	C	NC	C-NC
Share of experimental conditions																					
Control (0n/0p)	0.26	0.25	0.01	0.30	0.45	-0.15	0.18	0.21	-0.03	0.26	0.16	0.10	0.33	0.28	0.05	0.26	0.26	-0.00	0.26	0.25	0.01
Treatment 1 (1n/0p)	0.24	0.26	-0.02	0.30	0.13	-0.17	0.24	0.29	-0.05	0.32	0.30	0.02	0.14	0.32	-0.19	0.25	0.25	-0.00	0.24	0.21	0.03
Treatment 2 (1n/1p)	0.24	0.23	0.01	0.10	0.08	0.02	0.32	0.20	0.12	0.16	0.27	-0.11	0.30	0.17	0.13	0.24	0.23	0.02	0.21	0.28	-0.07
Treatment 3 (3n/0p)	0.26	0.26	-0.00	0.30	0.33	-0.03	0.26	0.30	-0.04	0.26	0.27	-0.00	0.23	0.22	0.01	0.24	0.26	-0.01	0.29	0.26	0.03
Demographics																					
Year of birth	1962.36	1961.81	0.56	1962.8	1956.85	5.95	1962.11	1961.26	0.85	1962.84	1962.55	0.29	1962.63	1962.50	0.13	1962.30	1962.20	0.09	1962.05	1961.47	0.59
Year of employment	1996.16	1996.31	-0.15	1993.9	1993.92	-0.02	1994.05	1991.05	3.00	1998.39	1997.65	0.74	1998.85	1999.80	-0.96	1995.47	1997.26	-1.79	1995.69	1995.36	0.33
Share of FTE≥0.9	0.80	0.78	0.02	0.80	1.00	-0.20	0.83	0.79	0.04	0.89	0.88	0.02	0.61	0.55	0.06	0.79	0.76	0.03	0.86	0.82	0.04
Share of female drivers	0.09	0.07	0.02	0.00	0.00	0.00	0.13	0.04	0.10	0.03	0.00	0.03	0.18	0.17	0.01	0.07	0.06	0.01	0.12	0.12	-0.00
Trip-specific variables																					
Punctuality	-2.94	-3.02	0.08	-4.56	4.57	0.01	-3.00	-2.87	-0.13	-2.68	-3.02	0.34	-2.59	-2.79	0.20	-3.11	-3.10	-0.01	-2.77	-2.87	0.10
Distance traveled	30.56	32.02	-1.46*	63.63	76.45	-12.82	32.49	31.53	0.95	32.75	30.52	2.23*	44.23	46.56	-2.34	20.69	22.27	-1.59	28.47	31.08	-2.61***
Number of passengers	15.03	15.28	-0.25	22.24	23.13	-0.90	15.24	14.79	0.46	15.47	15.38	0.08	18.97	22.03	-3.06**	11.96	11.18	0.78	15.51	16.05	-0.54
Number of bus stops	37.93	37.86	0.07	50.12	48.67	1.45	40.33	41.48	-1.15	47.36	42.14	5.22***	38.73	40.82	-2.09	31.54	31.15	0.39	37.24	39.20	-1.96
Share of rides:																					
- Morning rush hours	0.30	0.16	0.14***	0.41	0.04	0.37	0.19	0.19	0.01	0.27	0.23	0.04	0.25	0.17	0.08	0.32	0.14	0.18**	0.35	0.14	0.21*
- Evening rush hours	0.13	0.25	-0.13***	0.00	0.18	-0.18	0.19	0.22	-0.03	0.20	0.26	-0.06	0.17	0.22	-0.05	0.11	0.26	-0.15*	0.05	0.28	-0.24**
- Weekend	0.03	0.03	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.00	0.00	0.00	0.11	0.11	0.00	0.23	0.23	0.00	0.00	0.00	0.00
- fill-in	0.00	0.02	-0.02	0.00	0.00	0.00	0.00	0.03	-0.03	0.00	0.01	-0.01	0.03	0.07	-0.04	0.00	0.02	-0.02	0.00	0.03	-0.02
- Holidays	0.12	0.12	-0.01	0.00	0.00	0.00	0.20	0.21	-0.00	0.15	0.14	0.01	0.08	0.08	-0.00	0.09	0.10	-0.01	0.14	0.15	-0.01
- Urban area	0.15	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.44	0.01	0.00	0.00	-0.00
- School	0.004	0.004	-0.00	0.00	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.01	-0.00	0.00	0.00	-0.00
Share of rides on bus types																					
Bus type 1	0.77	0.76	0.02	0.55	0.38	0.17	0.93	0.96	-0.04	1.00	0.99	0.01	0.67	0.64	0.03	0.53	0.51	0.02	0.99	0.98	0.01
Bus type 2	0.08	0.10	-0.02	0.45	0.62	-0.17	0.07	0.04	0.04	0.00	0.01	-0.01	0.33	0.36	-0.03	0.02	0.05	-0.03	0.01	0.02	-0.01
Bus type 3	0.15	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.45	0.44	0.01	0.00	0.00	-0.00

Notes: coach logs are used to determine for every coaching date the mean difference in non-performance-related variables between drivers who received their first coaching (C) versus uncoached colleagues (NC) who were also working from the same base location on that date. Punctuality is the difference in minutes between actual and planned driving time. Distance traveled is measured in kilometers. Number of passengers is based on check-ins with public transport cards. Bus types 1 and 2 have diesel engines and bus type 3 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. Fill-in trips are trips in which a driver replaces a colleague from another base location. This table reports the mean over all coaching dates per base location. Stars indicate that the mean difference is significantly different from zero. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A3: Targeted Peer-Comparison Feedback Effects on Driving Performance: Other Driving Dimensions

	Braking					**Cornering**				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Post-announcement	-2.415***	-2.414***				0.654***	0.654***			
	(0.215)	(0.215)				(0.146)	(0.146)			
Post-feedback	-1.567***	-1.566***				-1.107***	-1.107***			
	(0.425)	(0.425)				(0.177)	(0.177)			
T1 (1n/0p)	-0.083					0.150				
	(0.554)					(0.340)				
T2 (1n/1p)	-0.555					-0.278				
	(0.498)					(0.309)				
T3 (3n/0p)	0.265					0.139				
	(0.540)					(0.320)				
Post-feedback × T1 (1n/0p)	0.253	0.252	0.029	-0.016	0.043	-0.037	-0.035	-0.082	-1.001	0.062
	(0.663)	(0.664)	(0.153)	(0.592)	(0.047)	(0.205)	(0.205)	(0.144)	(0.977)	(0.066)
Post-feedback × T2 (1n/1p)	0.415	0.414	0.059	0.189	0.011	-0.011	-0.010	-0.081	-1.237	0.110*
	(0.625)	(0.626)	(0.151)	(0.563)	(0.050)	(0.222)	(0.222)	(0.173)	(1.118)	(0.063)
Post-feedback × T3 (3n/0p)	-0.376	-0.378	-0.013	-0.036	-0.016	-0.180	-0.179	-0.090	-0.872	0.052
	(0.664)	(0.664)	(0.151)	(0.508)	(0.048)	(0.216)	(0.217)	(0.158)	(0.939)	(0.062)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Day fixed effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Trajectories	All	All	All	Urban	Rural	All	All	All	Urban	Rural
Number of drivers	398	398	398	120	389	398	398	398	106	389
Number of trip-level observations	473,042	473,042	473,042	135,337	337,705	462,170	462,170	462,170	118,263	343,907

Notes: identification of the treatment effects on driving performance. Standard errors in parentheses. The time period under consideration is until 15 November 2016, when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. 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. Post-announcement is a dummy variable with value 1 in the period after 1 November 2015 (feedback announcement), 0 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 after 15 December 2015. The dependent variables are braking and cornering, which are measured as the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, trips that were driven in a smaller or larger bus, and daily weather conditions (average temperature/wind and total rainfall). A no-report indicator is also included and captures drivers operating after 15 December 2015 (first feedback round) but who have not yet received their first report. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A4: Driving Performance Before and After the Experiment with Targeted Peer-Comparison Messages

	Fuel economy					**Acceleration**				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Post-announcement	0.097***	0.096***				-1.317***	-1.324***			
	(0.006)	(0.006)				(0.198)	(0.199)			
Post-experiment	-0.122***	-0.122***				0.202	0.206			
	(0.013)	(0.013)				(0.269)	(0.269)			
T1 (1n/0p)	0.016					-0.099				
	(0.023)					(0.274)				
T2 (1n/1p)	0.010					0.122				
	(0.026)					(0.312)				
T3 (3n/0p)	0.009					0.304				
	(0.023)					(0.267)				
Post-experiment × T1 (1n/0p)	0.004	0.004	0.002	n.a.	-0.000	0.218	0.218	0.281	1.083	0.050
	(0.018)	(0.018)	(0.017)		(0.017)	(0.363)	(0.363)	(0.362)	(1.021)	(0.252)
Post-experiment × T2 (1n/1p)	-0.006	-0.006	-0.006	n.a.	-0.006	-0.410	-0.423	-0.519	-1.739	0.060
	(0.016)	(0.016)	(0.016)		(0.016)	(0.533)	(0.533)	(0.527)	(1.692)	(0.198)
Post-experiment × T3 (3n/0p)	-0.000	-0.000	-0.003	n.a.	0.003	-0.706	-0.706	-0.686	-1.439	-0.190
	(0.018)	(0.018)	(0.017)		(0.017)	(0.559)	(0.559)	(0.514)	(1.412)	(0.239)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Day fixed effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Trajectories	All	All	All	Urban	Rural	All	All	All	Urban	Rural
Number of drivers	405	405	405	234	387	399	399	399	100	367
Number of trip-level observations	288,881	288,881	288,881	7,683	281,198	168,992	168,992	168,992	49,782	119,210

Notes: this table compares the pre-experimental performance of bus drivers to the post-experimental performance. Standard errors in parentheses. The pre-experimental period is the period before receiving the first feedback report. For most drivers, this was from 1 January 2015 until 15 December 2015. The post-experimental period starts after 15 November 2016 when it when it was communicated to the treated drivers that they will no longer receive peer-comparison messages. It continues until 31 January 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. Post-announcement is a dummy variable with value 1 in the period after 1 November 2015 (feedback announcement), 0 otherwise. The dependent variables are fuel economy (km/liters) and for acceleration it is the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, trips that were driven in a smaller or larger bus, and daily weather conditions (average temperature/wind and total rainfall). A no-notification indicator is also included and captures drivers operating in the post-experimental period but who have not received a feedback report around 15 November 2016 (and were thus not exposed to the notification message). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A5: On-The-Road Coaching Effects on Driving Performance: Other Driving Dimensions

	Braking						**Cornering**					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Post-coaching	-0.064	-0.299	-0.014				-0.048	-0.262	-0.047			
	(0.079)	(0.409)	(0.025)				(0.084)	(0.744)	(0.036)			
Day of first coaching				-0.594***	-1.776**	-0.108***				-0.787***	-2.648***	-0.185***
				(0.198)	(0.678)	(0.035)				(0.199)	(0.856)	(0.037)
Days after first coaching:												
1-7				-0.049	-0.224	-0.029				-0.111	-0.445	-0.104***
				(0.115)	(0.451)	(0.025)				(0.100)	(0.590)	(0.033)
8-14				0.093	0.327	-0.058				0.012	-0.085	-0.026
				(0.145)	(0.574)	(0.027)				(0.113)	(0.724)	(0.033)
15-21				-0.025	-0.204	-0.022				0.090	0.356	-0.083**
				(0.116)	(0.522)	(0.025)				(0.154)	(0.860)	(0.035)
22-28				0.083	0.141	-0.030				-0.005	0.049	-0.100***
				(0.145)	(0.559)	(0.026)				(0.158)	(0.902)	(0.037)
> 28				-0.099	-0.515	0.009				-0.023	-0.196	-0.013
				(0.090)	(0.475)	(0.030)				(0.118)	(0.972)	(0.044)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Trajectories	All	Urban	Rural	All	Urban	Rural	All	Urban	Rural	All	Urban	Rural
Number of drivers	391	112	376	391	112	376	391	98	376	391	98	376
Number of trip-level observations	262,332	75,205	187,127	262,332	75,205	187,127	251,829	58,605	193,224	251,829	58,605	193,224

Notes: identification of on-the-road coaching effects on driving performance. Standard errors in parentheses. The time period under consideration is the period for which we have complete logs available from all coaches (1 January 2015 - 30 April 2016). Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. The dependent variables are braking and cornering, which are measured as the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for post-coaching rides after an additional coaching session, travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, and trips that were driven in a smaller or larger bus. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table A6: Interaction Effects Targeted Peer-Comparison Feedback and On-The-Road Coaching: Other Driving Dimensions

	Braking			**Cornering**		
	(1)	(2)	(3)	(1)	(2)	(3)
Post-feedback × T1 (1n/0p)	-0.312**	-2.356**	-0.057	-0.208	-2.990***	0.017
	(0.157)	(1.182)	(0.059)	(0.143)	(1.041)	(0.091)
Post-feedback × T2 (1n/1p)	-0.112	0.875	-0.078	0.058	0.650	0.101
	(0.125)	(1.076)	(0.066)	(0.140)	(0.956)	(0.095)
Post-feedback × T3 (3n/0p)	-0.227*	-1.090	-0.138**	-0.180	-1.736**	-0.077
	(0.117)	(1.067)	(0.061)	(0.139)	(0.758)	(0.109)
Post-coaching × Post-feedback	-0.348**	-1.746	-0.114**	-0.050	-1.454**	-0.059
	(0.153)	(1.081)	(0.048)	(0.158)	(0.729)	(0.074)
Post-coaching × Post-feedback x T1 (1n/0p)	0.666***	3.278**	0.153**	0.218	2.665**	0.066
	(0.238)	(1.276)	(0.066)	(0.232)	(1.255)	(0.107)
Post-coaching × Post-feedback x T2 (1n/1p)	0.418*	-0.118	0.176**	-0.119	-1.283*	0.029
	(0.220)	(1.161)	(0.072)	(0.252)	(0.741)	(0.101)
Post-coaching × Post-feedback x T3 (3n/0p)	0.388**	1.444	0.214***	0.152	1.309	0.177
	(0.190)	(1.126)	(0.069)	(0.209)	(0.895)	(0.109)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Driver fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Bus type × day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Trajectories	All	Urban	Rural	All	Urban	Rural
Number of drivers	391	112	376	391	98	376
Number of trip-level observations	262,332	75,205	187,127	251,829	58,605	193,224

Notes: identification of the interaction effects between the two feedback programs. Standard errors in parentheses. The time period under consideration is the period for which we have complete coach logs available (1 January 2015 - 30 April 2016). Post-coaching identifies the day of first coaching and the full period thereafter for each coached driver. 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 after 15 December 2015. The dependent variables are braking and cornering, which are measured as the number of events per 10 kilometers (less events means better driving behavior). Regressions are estimated with robust standard errors (clustered by drivers) and control for travel distance, punctuality (difference between actual and planned driving time), number of passengers and bus stops, morning and evening rush hours, trips in urban areas, fill-in rides, and trips that were driven in a smaller or larger bus. A no-report indicator is also included and captures drivers operating after 15 December 2015 (first feedback round) but who have not yet received their first report. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.