university of groningen / faculty of economics and business

2024002-EEF

The Short and Long Term Effects of In-Person Performance Feedback: Evidence from a Large Bus Driver Coaching Program

Adriaan R. Soetevent
Gert-Jan Romensen

FEBRI is the research institute of the Faculty of Economics & Business at the University of Groningen. SOM has seven programmes:

- Accounting
- Economics, Econometrics and Finance
- Global Economics & Management
- Innovation & Organization
- Marketing
- Operations Management & Operations Research
- Organizational Behaviour

FEB Research Institute (FEBRI)
Faculty of Economics & Business
University of Groningen

Visiting address:
Nettelbosje 2
9747 AE Groningen
The Netherlands

Postal address:
P.O. Box 800
9700 AV   Groningen
The Netherlands

T +31 50 363 9090/7068

www.rug.nl/research/research-feb/

university of groningen

faculty of economics and business

The Short and Long Term Effects of In-Person Performance Feedback: Evidence from a Large Bus Driver Coaching Program

Adriaan R. Soetevent
University of Groningen, Faculty of Economics and Business, Department of Economics, Econometrics & Finance
a.r.soetevent@rug.nl

Gert-Jan Romensen
University of Groningen, Faculty of Economics and Business, Department of Economics, Econometrics & Finance

The Short and Long Term Effects of In-Person Performance Feedback: Evidence from a Large Bus Driver Coaching Program

Adriaan R. Soetevent*
University of Groningen

Gert-Jan Romensen†
University of Groningen

April 12, 2024

Abstract

Does in-person coaching structurally increases worker productivity? We answer this question using detailed performance data from a large-scale bus driver coaching program, exploiting natural variation in when drivers receive coaching.

Our results show significant improvements in fuel economy and outcomes pertaining to passenger comfort. Drivers with lower pre-coaching productivity experience the largest treatment effect, initially closing about 40% of the productivity gap. These effects last four to nine weeks and replicate in a control region. The data suggest that the main mechanism behind the coaching effect is information transmission from higher- to lower-productivity workers, rather than peer pressure or signaling.

JEL classification: D23, J24, M53, Q55.
Keywords: labor productivity, feedback, natural experiment, two-way fixed effects, peer learning

*Corresponding author. University of Groningen, EEF, P. O. Box 800, 9700 AV Groningen, The Netherlands, a.r.soetevent@rug.nl. We thank Viola Angelini, Robert Dur, Mikhail Freer, Marco Haan, Patrick Legros, John List, Erzo Luttmer, Erin Mansur, Noemi Pace, Noémi Péter, Peter Siminski and seminar participants at AFE 2016, BIOECON 2017, M-BEES & M-BEPS 2017, EEA 2017, IMEBESS 2017, ESA Vienna 2017, IAEE 2018, ESA Los Angeles 2019, EALE 2021, the workshop Recognition and Feedback at Erasmus University, the Arne Ryde Workshop, University of Adelaide, Universitat Autònoma de Barcelona, Dartmouth College, ECARES, Jiatong University Beijing, Maastricht University, Northeastern University, University of St.Gallen, 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. An earlier version of the paper has circulated under the title "Tailored Feedback and Worker Green Behavior". This study is registered in the AEA RCT Registry (#AEARCTR-0005391).

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

1 Introduction

Giving effective performance feedback is critical in maintaining and enhancing worker productivity, especially in work environments that hinder the use of pay-for-performance schemes (Blader et al. 2020, Gosnell et al. 2020). The adoption of digital monitoring technologies in the workplace has greatly expanded managers' scope to measure workers' performance and for giving workers tailored feedback (Staats et al. 2017).[1] These technologies routinely record additional data on the constituent parts of worker-level productivity, allowing for a precise measurement of the short- and long-term effectiveness of non-monetary incentives on different performance dimensions. This study contributes to the literature on the role of non-monetary incentives on worker productivity using detailed performance data from a coaching program launched by a major public transport company that is installing electronic on-board recorders (EOBRs) in its bus fleet. For every worker, EOBRs measure the trip-level aggregate productivity (fuel efficiency) as well as the underlying performance on acceleration, braking, and cornering (ABC).[2]

Following the EOBR installation, the company launched its EcoManager campaign to stimulate efficient and comfortable driving. As part of this campaign, the company introduced an in-person coaching program by experienced peers and in parallel to this monthly written individual feedback reports. In this study, we focuses on the effects of the in-person coaching program on drivers' performance. In-person coaching is an important way of tailoring feedback to disaggregate productivity measures. Recent studies have stressed the importance of peer-based learning for knowledge flows and skills transfers between workers (Sandvik et al. 2020, Lindquist et al. 2017, Chan et al. 2014). In the program that we examine, designated experienced drivers give individual-level coaching to their colleagues by riding along with them for a portion of their shift. At the end of the ride, the coach evaluates the trip with the driver and gives tips for improvement.

The timing of coaching can be viewed as the outcome of a quasi-random assignment process, as the coaches take a hop-on, hop-off approach to coaching and they have no access, by regulation, to the driver's performance: coaches select which drivers they will coach on a given day without taking into account a driver's past performance. We exploit the staggered adoption design

[1] Recent studies that examine how the adoption of electronic monitoring technologies by firms has impacted worker productivity include Pierce et al. (2015) and Kelley et al. (2018).

[2] In the first non-experimental studies that consider the EOBR technology, Baker and Hubbard (2003) and Hubbard (2003) study how the adoption of on-board computers has influenced the decision to integrate or outsource trucking services. Relatedly, Soleymanian et al. (2019) uses automotive telematics to study the effect of usage-based insurance on driving behavior.

feature of the program to uncover the short- and long-run causal impact of coaching on driving behavior. In non-experimental studies, feedback eligibility and intensity are likely to correlate with workers' (relative) productivity outcomes, which biases estimates of feedback effectiveness that compare worker productivity just before and right after the worker has received feedback. The (quasi-)random assignment of drivers to coaching considered in this paper avoids such selection problems.

There is surprisingly little experimental evidence on the impact of training programs initiated by firms. De Grip and Sauermann (2012) note that typical challenges in estimating the returns of training programs are the non-random selection of workers into these programs and the absence of a suitable measure to compare individual-level worker productivity before and after the training. The detailed coach logs and trip-level data on driving performance combined with the quasi-random phase-in of the coaching program allows us to address both challenges. Sandvik et al. (2020) accomplish something similar in an intricate field experiment that considers how frictions in knowledge transmission impact workplace knowledge flows. Using experimental variation in the joint incentives and the structure of the meetings that coworkers have, they examine the effect on knowledge transfers and sales. Our research design is also related to Gosnell et al. (2020) who conduct a large-scale field experiment with 335 airline captains. Compared to a control group with captains who were aware that they were in an experiment but received no further feedback, captains in the three treatment groups in addition received a monthly feedback report (delivered at their home address) with details on the captain's productivity outcomes in the past month (treatment group 1), plus personalized targets (group 2), plus prosocial incentives in the form of donations to charity (group 3). Comparing the outcomes of captains in the control group before and after the introduction of performance monitoring, they find that simple monitoring induces important improvements in labor productivity. Captains in the treatments with personalized targets show a further increase in productivity. The introduction of prosocial incentives has no measurable additional effect.

Our study complements these studies in the following ways. First, we target another major mode of transport in the logistics industry with a considerable impact on the environment: road travel accounts for three-quarters of all transport emissions (aviation for 12%) and all transport together accounts for one-fifth of global CO_2 emissions. Second, in focusing on in-person coaching we implement a different feedback design that adds to the modest number of studies that evaluate the impact of training programs on labor productivity. Third, because the company implements the feedback program region by region, we can use concession areas (200km away) as a holdout external control group, similar to Sandvik et al. (2020). This allows us to estimate the effect of the announcement and introduction of monitoring on fuel consumption with a difference-in-differences approach. Fourth, the recent difference-in-differences literature (de Chaisemartin and D'Haultfœuille 2020, 2022, Callaway and Sant'Anna 2021, Goodman-Bacon 2021, Sun and Abraham 2021) has shown that the commonly used two-way fixed effects (TWFE) specification may produce misleading estimates when treatment effects are heterogenous between treatment groups or over time. In experiments with staggered adoption designs, the variation in treatment timing can lead TWFE-estimated lead and lag coefficients to be biased. Estimated effects for a certain period can be contaminated by effects from other periods, implying that these estimates cannot be used to test for parallel pretrends. Hence, next to the set of estimates obtained through TWFE, we present a second set of estimates on the dynamic treatment effects of coaching that are obtained through an alternative estimator (Sun and Abraham 2021) that is free of such contamination. Finally, we use an empirical Bayes procedure (Morris 1983, Chandra et al. 2016) to estimate the ex ante worker quality metrics, which we use to categorize drivers into low, medium, and high productivity workers. Unlike a more traditional fixed effects approach (Lazear et al. 2016), this procedure corrects for measurement errors by extracting the stable component of worker quality. There exists a considerable productivity gap between workers. The average driver needs 24.38 liters of fuel to drive 100km, but after correcting for differences in driving conditions, the 90th percentile driver needs 2.67 liters more liters of fuel than the 10th percentile driver; a difference of more than 10%.

Our main findings are as follows. First, we find no improvement in fuel consumption following the announcement and introduction of the written feedback reports. This is in contrast to previous field research that, based on before-and-after comparisons, finds that the mere announcement of performance monitoring increases worker productivity (Blanes i Vidal and Nossol 2011, Gosnell et al. 2020). We do not observe that drivers in the treatment region improve relative to drivers in the control region during the announcement and introduction of EcoManager and the feedback program. Second, we observe strong and immediate effects of in-person coaching by peers. On the day of coaching the fuel need reduces by 0.51 liters/100km (0.42 SD, p <0.001) and the number of acceleration events by 1.06 events/10km (0.44 SD, $p < 0.001$). For braking and cornering behavior, these effects are 0.08 (0.04 SD, $p = 0.018$) and 0.19 (0.08 SD, p <0.001) events/10km. The improvements due to coaching tend to persist with a smaller magnitude in the ensuing weeks but fade out after four to nine weeks. Importantly, the estimated dynamic treatment effects on all four outcome variables replicate when we apply the alternative estimator by Sun and Abraham (2021). Zooming in, we find heterogeneity in the treatment effect of coaching on performance. The largest treatment effect is observed for the group of drivers with the lowest pre-coaching productivity, both at the day of coaching as in subsequent weeks; temporarily decreasing the worker productivity gap by more than one-third.

Our findings suggest that in improving worker productivity and conservation efforts, it is beneficial to employ face-to-face interventions. Our conceptual framework distinguishes three channels through which this effect may operate: knowledge transfers, signaling effects, and peer match effects. Our estimates point to knowledge transfers as the main mechanism driving the observed effects. Programs that facilitate interactions between workers and high-achieving colleagues can induce the transfers of skills and knowledge that are needed to foster positive changes in worker behavior. Even when the improvements at the worker level are relatively small, the aggregate impact on economic and environmental outcomes can be large and meaningful.

Section 2 of this paper motivates the study and reviews related literature. Section 3 presents a conceptual framework. Section 4 describes the field setting and the methods of analysis of the study. The empirical analysis follows in Section 5. Section 6 discusses the results and concludes.

2 Motivation and Literature Review

2.1 In-Person Coaching

A firm's growth in worker productivity may be hampered by the fragmentation and unequal distribution of knowledge and skills among its employee base. One inexpensive way of addressing this concern is to facilitate the within-firm knowledge exchange between workers and high-achieving peers. While learning from others is long recognized as an important input to individual learning mechanisms (Arrow 1994), it remains surprisingly unclear how firm-level initiatives to facilitate such exchanges affect workers' subsequent performance. Previous studies suggest that social interactions in the workplace improve workers' productivity mainly through peer pressure (Mas and Moretti 2009, Falk and Ichino 2006).[3] Less studies focus on productivity gains due to spillovers of knowledge and skills. How workers adapt to direct teaching from

[3] Peer pressure may in part also explain the increases in productivity found among data entry workers (Kaur et al. 2015), fruit pickers (Bandiera et al. 2010), call center employees (Lindquist et al. 2017), and salespersons of a department store (Chan et al. 2014). See Cornelissen et al. (2017) for a further discussion.

high-achieving colleagues is therefore a by and large open question.

A few studies, mainly in the areas of research and education, have examined productivity gains due to knowledge spillovers between peers (Jackson and Bruegmann 2009, Azoulay et al. 2010, Waldinger 2012). In a notable field experiment, Papay et al. (2020) find that student outcomes improve if low-skilled teachers are paired with high-skilled peers for coaching purposes. In workplace settings, Lindquist et al. (2017) report evidence of knowledge spillovers among call center employees. Chan et al. (2014) conclude that peer-based learning based on transfers of skills and knowledge is more important for worker productivity than learning by doing. Sandvik et al. (2020) observe persistent productivity improvements from peer learning.

The public transport setting and the coaching program initiated by our field partner are ideal to identify and evaluate the impact of skill and knowledge transfers between workers. First, by design we know exactly the direction of the transfer and the timing of the interaction. Second, the nature of the job leaves little room for other forms of learning from colleagues (e.g. observational learning). Bus driving is also a task in which performance is directly linked to individual effort. Finally, on-board computers in buses generate rich and high-frequency productivity data, and the short- and long-run impact of peer coaching can therefore be precisely estimated. We hypothesize that participating in a coaching session will, on average, improve the productivity of drivers vis-à-vis non-coached colleagues because they learn from high-achieving peers. However, signaling and peer pressure effects offer alternative explanations for observed productivity increases following coaching.

2.2 Stimulating Conservation Efforts among Workers

Next to contributing to the empirical literature on optimal feedback design in operations management, our findings also address the broader societal challenge of how to combat unsustainable energy consumption practices. There has been much progress in our understanding of non-financial incentives in residential energy consumption, yet research on how these insights generalize to firms is scant (Gerarden et al. 2017).[4] Our work helps to fill this gap and is part of an emerging literature that looks at the workplace for evidence on the effect of non-financial

[4]Existing studies on non-financial incentive schemes in the residential sector stress the importance of feedback and social approval in reducing energy (Allcott 2011) and water (Ferraro and Price 2013) consumption, 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 (Holliday et al. 2019) which reinforces the need for detailed evaluations of non-financial incentives pertaining to energy efficiency and also raises the question how these findings generalize to workers.

incentives on conservation efforts (Gosnell et al. 2020). Given that firms increasingly record and store data on multiple dimensions of worker-level productivity, tailoring feedback by decomposing consumption into its underlying sources seems a viable and promising approach to creating novel data-driven designs of conservation incentives (Brynjolfsson and McElheran 2016). The setting of a transport company is apt as the transport sector takes a heavy toll on the environment, accounting for one-fifth of global primary energy use and one-quarter of energy-related carbon dioxide (CO_2) emissions (IEA 2012). Indeed, the International Council on Clean Transportation hails fuel-efficient driving as low-hanging fruit to improve conservation levels (ICCT 2013). We examine how this can be achieved in settings where drivers have no financial stake in saving fuel.[5]

3 Conceptual Framework

To sharpen our thoughts, we introduce a theoretical choice of effort model that describes the bus driver's problem in choosing how much effort e to invest in efficient and comfortable driving. We build upon the model developed in Lazear et al. (2016), to which we add a peer-pressure component, cf. Silver (2021). Suppose that driver d's problem is to maximize an expected surplus $S(\cdot)$, represented as

$$S(e) = B(e) - \frac{C(e)}{q(t,t';k,\gamma)} - N(e;\phi_{d,c}(t;t')), \text{ with } k \geq 1. \tag{1}$$

The first term $B(e)$ denotes a driver's benefits of driving comfortably, which accounts for the effect of good driving on a driver's long-term chances of promotion but also the impact on a driver's self-image and the reward of receiving compliments and positive feedback from satisfied passengers. We assume $B(e)$ to be increasing and concave in e, $B'(e) > 0$, $B''(e) < 0$. The second term, $C(e)/q(t,t';k,\gamma)$, denotes the worker quality adjusted cost of effort. Following Lazear et al. (2016), we assume this cost to be increasing and convex in e, $C'(e), C''(e) > 0$, and furthermore to depend on worker ability/quality, indexed by $q(t,t';k,\gamma) \equiv k + m(t;t')k^{\gamma}$ with $k \geq 1$ the baseline worker quality.

[5] Choudhary et al. (2022) study the benefits of feedback on driving performance with an experimental design that varies the nature of the relative performance messages sent to car drivers who have downloaded a smartphone telematics app. Their setting differs from ours in a number of ways: their car drivers do have a financial stake in saving fuel, the feedback messages they use report an aggregate driving safety score to drivers instead of the disaggregate ABC-components (that together make up the safety score through a proprietary algorithm), and while the app does inform drivers on fuel efficiency, their data does not contain information on fuel economy.

We add the term $m(t;t')k^{\gamma}$ to the original Lazear et al. (2016) model to account for the effects of coaching on worker quality. The parameter γ determines whether the effect of coaching on worker quality is homogenous across workers ($\gamma = 0$), increases ($\gamma > 0$), or helps to decrease ($\gamma < 0$) the quality gap between lower and higher quality workers. If the mechanism behind the effect of coaching is information transmission, we expect coaching to benefit the lower performers in particular ($\gamma < 0$). A competing possibility in line with Lazear et al. (2016) is that coaching signals to workers that the company invests in performance improvement. Lower performers may respond more when they interpret this signal as the program affecting job security.

We assume for drivers coached at time t' that the incremental effect of coaching at time t on their baseline worker quality k equals k^{γ} times:

$$m(t;t') = \begin{cases} 0 & \text{for} \quad t < t' \\ b & \text{for} \quad t = t' \\ \alpha^{t-t'} b \geq 0 & \text{for} \quad t > t', \text{ with } \alpha \geq 0. \end{cases} \tag{2}$$

Here $b \geq 0$ denotes the instantaneous impact of coaching on worker quality. We assume this effect to be the same for all driver-coach combinations. For $b = 0$, coaching has no effect on worker quality. The parameter $\alpha \in [0, 1]$ represents the persistence of the coaching effect: the larger α, the more the effect of coaching sticks. For $\alpha = 0$, the effect of coaching is only there when the coach is present.

The third and final term $N(\cdot)$ is the effort norm a driver experiences when working with coach c. The company selected experienced drivers for the role of coach, which means that drivers are coached by their peers. This opens the possibility that any measured performance effect of coaching is not caused by an actual increase in worker quality due to information transmission, but is the result of peer pressure. Other studies have shown that the effort norm that workers experience may depend on the productivity of a worker's co-workers (Herbst and Mas 2015) and their (hierarchical) relationship to these co-workers (Bandiera et al. 2010, Gneezy et al. 2003). In our setting, we hypothesize that the peer pressure from coaching will lead to a higher salience ϕ of the effort norm during the coaching session, but possibly also afterwards each time the worker and coach have overlapping shifts. Hence we assume that $\phi_{d,c}(t;t') = \phi^H$ whenever the driver and coach have overlapping shifts following coaching ($t \geq t'$), and $\phi^L \leq \phi^H$ otherwise.[6]

[6]Salience may be higher at the time t' of coaching than afterwards when the shifts overlap, i.e. $\phi_{d,c}(t';t') \equiv \phi^C \geq \phi^H \equiv \phi_{d,c}(t;t')$ for $t > t'$.

We furthermore assume that the negative effect of peer pressure on utility is decreasing in effort ($N'(e) < 0$) and that this marginal effect of effort is higher the more salient the effort norm, $\partial^2 N(e;\phi)/\partial\phi\partial e < 0$.[7]

As mentioned above, coaching may not only cause pressure because it exposes drivers to a highly productive co-worker, but also because workers interpret the offered training as a signal from the company that it cares about performance and that especially the job security of the least-productive workers is threatened. In both cases, we expect the work effort by lower performing workers to respond more to coaching. In practice it is generally hard to separately identify the information transmission effect, the signaling effect and the peer pressure effect of coaching, as all predict that lower performers effort reacts most to coaching. However, one can argue that the information transmission effect is more likely to 'stick' than the signalling effect and increase in norm salience as these depend more on the coach's presence and hence tend to be more short-lived.

This gives the following results:

Result 1 *When coaching increases worker quality ($b > 0$), coaching increases a driver's effort at the time of coaching t', and also at times $t > t'$ when the effect is persistent.*

Result 2 *When $\gamma > 0$ ($\gamma < 0$), coaching increases (decreases) the worker-quality gap; for $\gamma = 0$, the effect of coaching on worker quality is homogenous across workers.*

Result 3 *When coaching increases the salience ϕ of the effort norm, coaching at time t' increases a driver's effort at time t', and at time $t > t'$ only when the shifts of the driver and coach overlap.*

The first result is obtained by taking the first order condition for equation (1) and using the implicit function theorem:

$$\left.\frac{de}{dm}\right|_{FOC} = -\frac{\partial S/\partial m}{\partial S/\partial e} = -\frac{\partial^2 S/\partial m \partial e}{\partial^2 S/\partial e^2} > 0.$$

The relation between effort and coaching results because the numerator $\partial^2 S/\partial m\partial e = \frac{C'(e)k^{\gamma}}{(k+m(t;t')k^{\gamma})^2} > 0$ for $k \geq 1$ and because the denominator is the second derivative of the objective function which

[7] For example, if the maximum possible effort level is $\bar{e}$, the peer-pressure function might take the form $N(e;\phi) = \phi(\bar{e}-e)^2$, in which case $N'(e) = -2\phi(\bar{e}-e) < 0$ and $\partial^2 N(e;\phi)/\partial e\partial\phi = -2(\bar{e}-e) < 0$.

must be negative for the solution to be an interior maximum. Results 2 and 3 follow similarly, see Appendix A for the full proofs.

4 Field Setting and Research Design

4.1 Industry

Our field partner is Arriva, a European-wide passenger transport company operating various transport modes in public transport. Bus transport is the firm's largest business unit.[8] In the Netherlands, bus concessions are granted to companies by means of a tendering procedure. Winning a tender gives companies the exclusive right to operate in a designated area for a number of years. To stimulate firms to engage in environmentally friendly behavior and to improve the living conditions of its citizens, environmental objectives feature prominently in the requirements tendering parties need to meet.[9] In the long run this trend may drive bus companies to buy vehicles with a hybrid or electric fuel technology. In the short run, the installment of electronic on-board recorders (EOBRs) helps them to meticulously measure performance on several dimensions of driving behavior. For example, the version used by Arriva records trip-level performance on fuel consumption and comfort dimensions such as acceleration, braking and cornering (ABC). Each driver logs into the system with a unique identifier to match the performance records and trip-related background variables. This enables precise monitoring and generates a wealth of high-frequency data on worker productivity and conservation efforts.

The system calibrates the comfort dimensions as follows. Based on test rides under different circumstances, the company has formulated threshold performance levels for every dimension. Technically, the thresholds relate to minimum G-force measurements by a three-axis accelerometer in the bus. During each trip, the EOBR records an 'event' whenever an action by the driver is in excess of these thresholds. The performance measure of the ABC dimensions is the number of events per 10km, with fewer events indicating better driving. These outcome data are linked with centralized databases containing information on a host of driver and trip characteristics, providing a detailed picture of driver performance over time under various road conditions.

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

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

4.2 Research Setting and Timeline

As part of its EcoManager campaign, Arriva Netherlands installed new EOBRs in its entire fleet in the time period 2015-2017. The EOBR data are used as input to monthly feedback reports that are distributed among the drivers. In addition, the company introduced a coaching program in which drivers receive feedback and advice from a selected experienced colleague during on-the-road sessions. The new technology and the feedback programs are phased in over time in the different concession areas. We join the implementation process in the first concession area in the Fall of 2015. This area comprises about two-thirds of the province of Friesland in the Netherlands where Arriva serves about 5.16 million travelers in a year.[10] We refer to this area as the treatment region. In two other concession areas, located in the southwestern part of the Netherlands, the company implemented the feedback program at the end of 2016. This area serves as our control region.

Figure 1 shows the timeline of the study. We use the old on-board system to establish a long baseline of fuel consumption, starting in January 2015. The recording of fuel economy data in the control region starts in July 2015 which allows us to compare and contrast the fuel economy scores by drivers in the treatment region with those of drivers in the control region in the same period. In the treatment region, the new EOBR system provides baseline data on the ABC dimensions from September 2015 onwards. Unfortunately, unlike for fuel econonomy we cannot contrast these with ABC outcomes in the control region for the same period as this information is only recorded in the control region from September 2016 onwards.

In the baseline period, drivers are not informed about the upcoming feedback, nor that they are being monitored. The company sent promotion material about the EcoManager campaign to the different locations on October 5, 2015. The project was officially launched with a kickoff event on November 9, 2015. At this date, the LED-array in the buses is also switched on, providing drivers with some instant feedback.[11] At the event, all drivers were informed about the digital monitoring and the introduction of monthly individualized feedback reports starting in December 2015. We use the post-announcement period (09 Nov. - 15 Dec. 2015), the diagonally shaded area in Figure 1) to estimate the combined effect of the announcement and LED activation.

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

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

From December 15 onwards, drivers receive monthly feedback reports. The general report gives drivers feedback on their performance on the ABC comfort dimensions as well as fuel economy. They receive a letter score, ranging from A (highest score) to D (lowest score), which is determined relative to a general pre-set benchmark such that in principle all drivers can score As.[12] In addition, the field experiment we report on in the companion paper (Romensen and Soetevent 2024) integrates peer-comparison feedback messages into this general feedback report. The peer-comparison messages inform drivers how they rank against their peers rather than against an overall benchmark on the ABC comfort dimensions.[13] The peer-comparison feedback period ends November 15, 2016, when drivers receive a one-time notification that the peer-comparison messages are no longer included in the reports.[14] They then enter the post-experimental period.

Drivers were informed at the kickoff that the feedback will not be used in formal evaluations. This may rule out career concerns as an alternative explanation, but we note that it runs counter to the firm's objectives to follow through on this claim (Hölmstrom 1979). Apart from the feedback programs considered, no other incentives were used by the company to promote energy-efficient driving among workers.

4.3 Driver population, buses and routes

Most drivers in the treatment region are tenured employees, while a small minority of about 14% operates on a temporary contract. Drivers are typically not involved in other tasks within the organization. Opportunities for promotion are limited and the work council is against using financial incentives to reward good performances.[15] Before EcoManager, drivers did not receive personal feedback. Each driver belongs to one of six base locations. For five locations, virtually all routes are between cities and in rural areas. Routes are based on timetables and do not vary much over time. The sixth and largest location has a mixture of urban and rural routes. Within a location, drivers' weekly shifts rotate. This implies that the worker faces week-to-week

[12] A sample feedback report can be found in Appendix section E.

[13] There is treatment variation in feedback intensity: Each driver is randomly assigned a group that receives combinations of up to x corrective, and up to y positive messages that target the underlying ABC driving dimensions. The four treatment groups are T1 (Control): $x = 0, y = 0$; T2: $x = 1, y = 0$; T3: $x = 1, y = 1$; T4: $x = 3, y = 0$. See Romensen and Soetevent (2024) for further details on the treatment design and results.

[14] 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 included in the report that was distributed in November 2016 to all drivers that were in the treatment conditions with peer-comparison feedback (all drivers except those in the control condition).

[15] The design of conservation incentives within firms is often dictated by institutional constraints that hinder the use of pay-for-performance schemes.

variation in his or her assignment to trips and the schedule repeats after about 14 weeks. This way of scheduling ensures that drivers are familiar with their routes and drive each route under different on-the-road circumstances. The schedules generate within-location variation in the type of trips. An important effect of the shift rotation is that each route is assigned to multiple drivers. This allows us to include a rich set of fixed effects in our empirical analysis.

In the treatment region, the company operates three bus types: VDL, Intouro and IRIS. The VDL bus is most commonly used, accounting for 75% of all trips. Intouro buses are mainly used for longer distance routes. Other than the VDL and Intouro buses that run on diesel, the IRIS bus runs on natural gas. This implies that for trips completed with an IRIS bus, no fuel economy is reported. In addition, the IRIS bus is only used in one base location, whereas the VDL and Intouro buses are used in all locations. Hence, we use the sub-sample of trips completed with either a VDL or Intouro bus to estimate the treatment effects. We match the trip-level observations in the final sample with driver, trip, and daily weather characteristics.[16]

4.4 Natural Experiment: In-Person Coaching

The object of this study is the face-to-face coaching program that was initiated by the company. This coaching program starts around the kickoff event in November 2015. We ensured that the coaches kept detailed diaries on who was coached when. Most drivers receive their first coaching in the weeks following the kickoff event.[17] During this period, the company reserved extra time for the coaches to ride along with drivers and to answer questions related to the upcoming feedback. Coaching intensity gradually decreases until it levels off after the first feedback report in mid December 2015. We have complete coach logs for the period till April 30, 2016. Some coaches no longer provided or kept track of coaching after that date. For this reason, we focus on the period till April 30, 2016 (the light-shaded area in Figure 1) for which we have complete data availability on fuel economy and coaching activities for all drivers in the treatment and control region, plus all data on ABC-outcomes in the treatment region once the EOBR system had started to measure and register these data. At the end of this period, 110 drivers (27% of the full sample) were never coached and 21 drivers (5%) participated in additional coaching sessions.

Based on their driving track record, six experienced drivers (one per base location) were

[16] Appendix section F details the steps we have taken to construct the final sample.
[17] Appendix section D gives full details on the timing of coaching.

recruited to act as coaches. They received training on how to approach drivers and how to communicate feedback. The coaches are bus drivers themselves and their time for coaching activities is limited to about one day per two weeks.[18] The coaches take a hop-on hop-off approach to on-the-road coaching. This means that during the day, coaches must select the driver for their next coaching session from the group of drivers who happen to arrive at the bus station, around the time the previous session ended. This makes random assignment of coaching sessions at the driver-trip level impossible. At the same time, this and the fact that coaches have no access to the individual feedback reports ensures that coaches can hardly target specific drivers with poor scores. We will provide empirical support for the view that the assignment of drivers to coaching is the outcome of a quasi-random process.

In a coaching session, a coach rides along with a bus driver for a portion of the driver's shift. This allows the coach to tailor the feedback to driver-specific issues. 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 gives 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.[19] Drivers are treated as equals and feedback is delivered in a constructive and positive manner.

Coaches maintain a detailed log of their activities, allowing us to pinpoint when and how often drivers are coached. To check whether the assignment of coaching can be considered random and not based on pre-selected criteria, we compare for each outcome variable (fuel economy and ABC comfort) the mean baseline performance of drivers who have received their first coaching and non-coached colleagues with the same base location. Table 1 verifies balancing on multiple baseline outcome performance measures and covariates. We present both the standard p-values and the ones adjusted for the problem of multiple hypothesis testing using the Bonferroni and Holm correction. Only for morning and evening rush hours we find statistically significant differences which merely reflects that coaches tend to start early in the morning. For none of the other variables we find differences that are even close to significance. This supports the view that the implementation of the coaching program exhibits a quasi-random order of phase-in.

[18]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.

[19]The logs do not contain these notes, so we do not know exactly what the coach communicated.

4.5 Worker Quality Metrics

The framework presented in (1) and Result 2 emphasize that drivers may respond differently to the coaching treatment dependent on their underlying worker quality. Hence we need an estimate of a driver's pre-coaching performance. One approach is to consider each driver's pre-coaching trip level performance on the ABC and fuel economy outcome dimensions and to regress these on a rich set of explanatory and contextual variables to extract driver fixed effects which can serve as a worker quality metric. Lazear et al. (2016, Section 3.E) use this strategy to classify workers with an above (below) median fixed effect as "stars" ("laggards"). However, because for some drivers a smaller number of pre-coaching observations is available than for others and because the outcomes themselves have a stochastic component, the fixed effects may suffer from measurement error due to noise from sampling variation which can lead to attenuation bias in our analysis. To address this concern, we adopt a step-wise approach similar to Chandra et al. (2016) who study the effect of hospital quality on hospital market share.

First, to reduce the influence of stochastic month-to-month variability in performance we use the complete set of a driver's pre-coaching trip level observations to construct our pre-coaching worker quality metrics (for fuel economy and ABC). Second, in our analysis of treatment effect heterogeneity across the performance distribution, we restrict the sample to drivers with at least 25 pre-coaching trips.[20] Third, we correct for estimation error in the fixed effects using an empirical Bayes shrinkage procedure (Morris 1983). The issue is that when a driver's estimated fixed effect is at the far lower (upper) end of the distribution, it is likely to suffer from negative (positive) estimation error. The empirical Bayes (EB) procedure adjusts for this by extracting the stable ('true') component of worker quality. The EB-adjusted quality is given by

$$\mu_i^{EB} = (1 - b_i)\hat{\mu}_i + b_i \mathbf{z_i'}\gamma, \tag{3}$$

with $\hat{\mu}_i$ the estimated driver fixed effects that result from a trip level regression and $\mathbf{z_i'}\gamma$ the mean of the 'true' underlying worker quality. In the implementation, we assume the latter to equal a base location fixed effect. Finally, the weight $b_i \in [0, 1]$ is the attenuation factor or shrinkage coefficient: the higher the signal-to-noise ratio, the lower the value of b_i and the higher the weight given to the estimate. The EB-adjusted estimate of worker quality is a weighted average of the estimated fixed effect and the underlying quality process, where the weight given to the

[20] For fuel economy, we include 405 (99%) of all drivers, for the ABC outcomes 372 (91%). The somewhat lower number for the ABC-outcomes results from the shorter baseline period.

prior mean is inversely related to the precision with which the fixed effect is estimated. See Appendix B for details on this procedure.

Table 2 (Panel A) reports the average and median estimated attenuation factor for each of the worker quality metrics. For each metric, the median estimate is smaller than 0.01, the averages are also small but somewhat larger for braking and cornering. Overall, the estimated attenuation factors are close to zero which means that for most drivers the driver fixed effect is estimated with considerable precision. This is because we have on average 317 (ABC comfort dimensions) to 690 (fuel economy) pre-coaching trip level observations per driver.[21] We use the EB-adjusted fixed effect estimates ($\hat{\mu}_i^{EB}$) to establish for each outcome a driver's pre-coaching percentile rank $p(i)$. A lower rank corresponds to a better performance. Reassuringly, the employees who act as coaches belong to the best-performing drivers themselves. Panel C of Table 2 shows that on all metrics, all coaches belong to the top 50% of drivers; for fuel economy and acceleration they all are in the top 25% of drivers. Using the raw instead of the EB-adjusted fixed effect estimates would lead to a similar ranking of drivers. For all outcome dimensions, the correlation in rankings exceeds 0.98, reflecting the small estimated attenuation factors.

Scope for Improvement Given the estimates of the driving condition adjusted worker quality metrics and the standard deviations that account for measurement error, we are now ready to explore the scope for improvement that the coaching program might entail. Which factors influence driver performance on fuel economy and the ABC dimensions? What part of this performance is within a driver's sphere of influence and what part is caused by external factors such as weather and other driving conditions? Panel A of Table 2 reveals sizable between-driver variation in performance. Using the number of observations per driver as weights, we observe that the average driver needs 24.38 liters of fuel to drive 100km.[22] Part of this variation can be attributed to differences in driving conditions. Regression estimates (reported in Online Appendix Table C.1) show that the bus type accounts for 30.4 percent of the between-trip variation in fuel economy, with longer buses having a sizable and significantly worse fuel economy. The impact of weather conditions is limited. As one expects, fuel economy is negatively correlated with the number of stops per kilometer, the number of passengers, evening rush hours and the bus running late. These variables seem to capture most of the day-to-day variation in fuel

[21] To compare, across health conditions Chandra et al. (2016) have an average of 66 to 102 patients per hospital.

[22] 25 liters/100km $\sim$ 10.6 gallon/100miles. Throughout the text, we will state (changes in) fuel economy in l/km instead of km/l because of the miles-per-gallon (MPG) illusion (Larrick and Soll 2008).

economy, as adding day fixed effects only slightly improves the R^2. Structural differences in driver performance explain an additional eight percentage points of variation in trip-level fuel economy. After controlling for the trip-level driving condition, the variation in fuel economy performance between drivers as measured by the EB-adjusted standard deviation that accounts for measurement error is $\hat{\sigma}_\xi = 1.20$.[23] Hence, the potential for improvement is economically significant: A policy able to move a driver's average fuel economy from the 90th percentile to the 10th percentile reduces this driver's fuel bill by 2.67 liters/100km or about 10%.

We use the coefficient of variation $c_v = \hat{\sigma}_\xi/\bar{y}$, with $\bar{y}$ the mean of the outcome variable of interest as a standardized measure of dispersion to compare the relative scope for improvements in fuel economy and in the ABC dimensions, cf. Vivalt (2020). For fuel economy, this coefficient equals 0.049 (= 1.200/24.379). The coefficients of variation shown in Table 2 reveal that in relative terms, between-driver dispersion is larger for the ABC dimensions than for fuel economy. However, for braking and cornering the average number of events per 10km is relatively close to the zero lower bound, limiting the upward potential for a large fraction of drivers.

Of course, the different outcomes are related: more acceleration events for instance increase fuel consumption. Table 2 Panel B shows the correlation between fuel economy and the comfort dimensions after controlling for a rich set of trip-level characteristics.[24] Fuel economy is correlated with acceleration and, to a lesser extent, with braking and cornering. This implies that next to being worker productivity measures in their own right, improvements in either of them is likely to also contribute to fuel economy.

5 Results

5.1 Announcement and Introduction of General Feedback

Before we examine the effect of coaching on worker productivity, we first implement a standard difference-in-differences (DID) methodology to test whether the announcement and introduction of EcoManager and the general feedback program improved the fuel economy (in liters per 100km) of drivers in the treatment region relative to drivers in the control region. For each region, Figure 2 shows the development of the weekly average fuel economy over time. Tables 3 and 4 show crosstabs of the mean realized fuel economy. In each table, the data are split into four cells. The columns split the data by region: the treatment versus the control region. The

[23] Online Appendix C contains the regression estimates for all outcome dimensions.
[24] This set is the same set as in column (4) of Online Appendix Table C.1.

rows split the data by period: a baseline period and an intervention period. Each cell contains information on the mean fuel economy, together with the standard error and the number of trips. We cluster the standard errors by date to account for possible correlated shocks across trips (Bertrand et al. 2004).[25]

Table 3 considers the effect the announcement of EcoManager, i.e. the aggregate effect of the launch of the campaign and the switching on of the LED-arrays (which happen on the same date). The mean fuel economy of drivers in the treatment region shows a marginal increase of 0.032 liters/100km during the post-announcement period relative to the baseline period. At the same time, the mean fuel economy in the control region experiences a very similar increase of 0.144 liters/100km. Hence, relative to the control region where drivers did not receive an announcement, the fuel consumption by drivers in the treatment region declines by an insignificant $DID_{Ann} = -0.112$ liters/100km (s.e. 0.107). We cannot reject the null of no announcement effect. Thus we do not replicate previous research (Blanes i Vidal and Nossol 2011, Gosnell et al. 2020) that found that the mere announcement of performance monitoring increases worker productivity. One reason for this difference may be (unobserved) differences in (organizational) setting and elements of the monitoring program. Another possible reason is that these previous studies are based on before-and-after comparisons, whereas our DID-estimates use an external control group at 200km distance. When we do a before-and-after comparison on fuel consumption using data from the treatment region only and using the controls as in column (2) of Table 5, we find an announcement effect of -0.282 liters/100km (s.e. 0.073), significant at $p < 0.001$.

Table 4 looks at the effect of the distribution of the feedback reports, which starts in the treatment region on December 15, 2015. The rows split the data into a baseline period and a post-feedback period. The mean fuel economy of drivers in the treatment region drops by 0.500 liters/100km during the post-feedback period relative to baseline. However, the mean fuel economy in the control region also decreases by 0.428 liters/100km. Hence, the estimate $DID_{GenFB} = -0.072$ (s.e. 0.076) indicates that relative to the control region where drivers did not receive reports, the fuel consumption by drivers in the treatment region declines by a statistically insignificant 0.072 liters/100km. We cannot reject the null of no effect of the general feedback reports.

[25]We cluster by date instead of by driver because this leads to the most conservative (i.e. largest) standard errors (Cameron and Miller 2015). Abadie et al. (2023) show that these standard errors can be too large. In our case, standard errors are similar when we cluster instead at the driver level.

In the treatment region, drivers start to receive coaching following the kickoff event. The feedback reports of drivers in the treatment groups contain peer comparison messages next to the general feedback (see Figure 1). The DID-estimates do not account for these changes and hence we evaluate the robustness of our estimates by estimating the following regression specification that includes a rich set of covariates.

$$Y_{its} = \mu_i + \lambda_t + \gamma \mathrm{PF}_{it} + X_{its} \cdot \theta + \sum_{j=2}^{4} \tau_j \mathrm{PF}_{it} \times \mathbf{1}(T_i = j) + \kappa_b + \zeta_{bt} + \upsilon_{its}. \quad (4)$$

The outcome variable of interest (fuel economy or ABC) is denoted by Y_{its}, indexed by driver (i), time in days (t), and the bus trip (s). The binary variable PF_{it} (post-feedback) equals one once driver i has received the first feedback report and zero otherwise. This definition makes no selection on the actual reading of the report. This is useful from a policy perspective because it captures the aggregate performance of the treatments when applied to an eligible population (Allcott 2011).[26] The vector X_{its} contains additional covariates (listed in Table C.1). Importantly, because coaching takes place in parallel to feedback, post-coaching dummy variables are included in the controls. The τ-coefficients correct for the treatment-specific effects of receiving peer-comparison feedback messages with $\mathbf{1}(T_i = j)$ indicating whether driver i in the treatment region is assigned peer-comparison treatment j, $j = 2, \ldots, 4$. A rich set of variables controlling for driver (μ_i), bus type (κ_b), day (λ_t), and bus type interacted with day (ζ_{bt}) fixed effects completes the specification.[27]

In this specification, the γ-coefficient estimates the effect of the introduction of general feedback. Panel A of Table 5 shows that for fuel economy, $\hat{\gamma} = 0.002$ (s.e. 0.077) once we control for day fixed effects. Similarly, when we replace PF_{it} in (4) by PA_{it} (a variable that equals one if the observation is in the post-announcement period), γ estimates the aggregate effect of the EcoManager launch. We find $\hat{\gamma} = -0.021$ (s.e. 0.055). These estimates are essentially identical to our insignificant difference-in-differences estimates.[28] As we report in Romensen

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

[27]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 of avoiding ABC events due to bus type specific threshold settings.

[28]To separate the effects of the announcement and introduction of EcoManager from the effect of coaching, we also computed the DID-estimates while only including drivers in the treatment region who did not receive coaching. The estimates, $DID_{Ann} = -0.036$ (s.e. 0.114) and $DID_{GenFB} = -0.113$ (s.e. 0.079), are similar to the ones reported.

and Soetevent (2023), the point estimates of the τ-coefficients are small in size and individually and jointly insignificant.

In sum, a DID approach that compares the change in mean fuel economy in the treatment region (Friesland) with the change in fuel economy in the control region (South Holland) over the same time span finds no indication that either the announcement or the introduction of the general feedback program has improved worker productivity. We cannot use DID or regression models that include day fixed effects to run a similar analysis for the ABC productivity measures that capture driving comfort, because we lack ABC outcome data for the control region for the period congruent with the post-feedback period in the treatment region. The estimates without day fixed effects in Table 5 Panel B suggest that general feedback importantly improves performance on all comfort dimensions. Yet, we refrain from giving meaning to these estimates because they are based on before-and-after comparisons and our earlier results show that for fuel economy these turn out not to be robust.

5.2 In-person Coaching

To identify the overall effect of a single on-the-road coaching session on productivity outcomes in the weeks following, we estimate both a static and a dynamic two-way fixed effects (TWFE) specification. The static specification estimates a time invariant treatment effect of coaching. The dynamic specification allows the effect of coaching to vary over time in a non-parametric fashion. Let t_i^c denote the specific calendar day at which driver i receives coaching and define $D_{i,t}^{post} \equiv \mathbf{1}[t > t_i^c]$ to be an indicator of driver i having received the initial coaching session by time t. For never-treated drivers (27% of all drivers in our sample), $D_{i,t}^{post} = 0$ for all t. We estimate the following static specification:

$$Y_{its} \quad = \quad \mu_i + \lambda_t + \delta^{post} D_{i,t}^{post} + X_{its} \cdot \theta + \kappa_b + \zeta_{bt} + \epsilon_{its}. \tag{5}$$

In this and the dynamic specification that follows, the dependent variable Y_{its} denotes as before the outcome of interest, the same set of control variables as in equation (4) is included and standard errors are clustered at the driver level. The estimate of δ^{post} captures in a single number the effect of coaching in the entire post-coaching period.

We also estimate the following regression specification with dynamic treatment effects which allows the effect of coaching to vary non-parametrically over time:

$$
\begin{aligned}
Y_{its} &= \mu_i + \lambda_t + \sum_{\tau=-10}^{-2} \delta_\tau D_{i,t}^\tau + \delta_0 D_{i,t}^0 + \sum_{\tau=0}^{10} \delta_\tau D_{i,t}^\tau + \sum_{\tau=\{-,+\}} \delta^\tau D_{i,t}^\tau \\
&+ X_{its} \cdot \theta + \kappa_b + \zeta_{bt} + \epsilon_{its}.
\end{aligned} \tag{6}
$$

In this specification, $D_{i,t}^0 \equiv \mathbf{1}[t = t_i^c]$ is an indicator function for driver i being coached at day t_i^c. Similarly, the indicator functions $D_{i,t}^\tau \equiv \mathbf{1}[7(\tau-1) < t - t_i^c \leq 7\tau]$ for $\tau = 1, 2, \ldots, 10$ ($D_{i,t}^\tau \equiv \mathbf{1}[7(\tau) \leq t - t_i^c < 7(\tau+1)]$ for $\tau = -10, -9, \ldots, -1$) each denote a seven-day period after (before) coaching and together cover a total of twenty weeks. For example, $D_{i,t}^2 = 1$ from day eight to day fourteen after driver i has received coaching. We follow common procedures (Sun and Abraham 2021) and bin the more distant relative periods by defining $D_{i,t}^- \equiv \mathbf{1}[t - t_i^c < -70]$ and $D_{i,t}^+ \equiv \mathbf{1}[t - t_i^c > 70]$. The estimate of δ_0 captures the treatment effect of coaching at the day of coaching and the treatment lag (lead) coefficients $\delta_1, \ldots, \delta_{10}$ ($\delta_{-10}, \ldots, \delta_{-2}$) the dynamic effects in the first ten weeks following (preceding) coaching. The coefficient δ^+ absorbs any impact of coaching after 10 weeks. The coefficient δ^- pools the effects in the period more than 10 weeks before coaching. Period -1 (the seven days before coaching) is the omitted period.[29]

Figure 3 and the odd columns in Table 6 show the dynamic treatment effects of coaching for the fuel economy and ABC outcomes. For fuel economy and acceleration, we observe a strong and immediate effect of coaching: The fuel need reduces by 0.51 liters/100km ($0.42\hat{\sigma}_\xi$) on the day of coaching and the number of acceleration events by 1.06 events/10km ($0.44\hat{\sigma}_\xi$), $p < 0.001$ in both cases. For cornering and especially breaking the effects of coaching are less pronounced. For these two outcomes, the baseline number of events per 10km is already relatively close to the zero lower bound. For both, the effect of the day of coaching is still significant but smaller: -0.19 events/10km ($0.08\hat{\sigma}_\xi$, $p < 0.001$) for cornering and -0.08 ($0.04\hat{\sigma}_\xi$, $p = 0.018$) for braking. For all outcomes except braking, the effects of coaching persist for four to nine weeks but clearly decay as time progresses.

With reference to the conceptual framework we find that coaching seems to increase worker quality ($b > 0$, Result 1) and that this effect sticks, but only temporarily ($0 < \alpha < 1$). The result that coaching does leave a temporary but clearly measurable effect after the coach has left is a first indication that the main mechanism behind coaching is information transmission rather than signaling. Drivers seem to fall back into old driving habits, fitting evidence from many

[29]The estimated δ-coefficients in Figure 3 show the average effect relative to this baseline period.

domains that it is hard to induce persistent changes in habits (Brandon et al. 2019). For none of the outcomes we observe differences in driving behavior in the ten weeks prior to coaching, which lends support to our earlier conclusion that the selection into coaching is quasi-random and not based on prior performance.

Recently, there has been considerable debate on the use of the TWFE estimator in staggered adoption designs (de Chaisemartin and D'Haultfœuille 2020, 2022, Callaway and Sant'Anna 2021, Goodman-Bacon 2021, Sun and Abraham 2021, Sun and Shapiro 2022, Athey and Imbens 2022, Wooldridge 2021, Baker et al. 2022). Sun and Abraham (2021) point out that in settings like ours with variation in treatment timing, the coefficients on leads and lags in TWFE specifications like (6) can be contaminated by effects from other periods, especially when there is heterogeneity in how drivers coached at different times respond to coaching. Such heterogeneity for example arises if coaches select the initial treatment timing of a driver based on expected treatment effects, instructing the worse (or better) drivers first. The balance tests in Table 1 do not suggest that this is a source of heterogeneity in our setting, yet one can easily think of other causes. For example, time variation in weather or road conditions may cause one cohort of drivers to experience a treatment path that is different from drivers treated at a different time. An implication is that in the presence of heterogeneity one cannot use the estimates of the lead coefficients $\delta_{-10}, \ldots, \delta_{-1}$ to test for parallel pretrends.

To account for this and as a robustness check, we re-estimate equation (6) using the interaction-weighted (IW) estimator developed by Sun and Abraham (2021). Simulation evidence in Baker et al. (2022) shows that this estimator is effective for estimating treatment effects in settings with variation in treatment timing and heterogeneous treatment effects. This alternative estimator uses the cohort average treatment effect $CATT_{e,\ell}$ as its basic building block. $CATT_{e,\ell}$ represents the average treatment effect ℓ periods past the initial treatment for the cohort of units that were first treated at time e. In estimating each $CATT_{e,\ell}$, the group of never-treated drivers in the sample constitutes the control group. The resulting IW estimator $\hat{\nu}_\ell$ is a weighted average of the $CATT_{e,\ell}$ estimates, with the weights equal to the share of each cohort in period ℓ. Other than the conventional TWFE estimates δ_t and γ_t, the IW estimates $\hat{\nu}_\ell$ are guaranteed to fall within the convex hull of the underlying $CATT_{e,\ell}$ estimates and are not contaminated by the $CATT_{e,\ell'}$ estimates from periods $\ell' \neq \ell$.

To apply the IW estimator to our setting, we use calendar weeks as periods and define a cohort as the group of drivers that received coaching in the same week, ℓ is the number of calendar weeks since the initial coaching session. We decided not to include any other explanatory variables except for the time and driver fixed effects because Sun and Abraham (2021) establish the validity of the IW estimators for panel data without covariates.[30] Figure 4 and the even columns in Table 6 present the results. A comparison with the previous $\hat{\delta}$-estimates leads to a number of observations. First, $|\hat{\nu}_0| < |\hat{\delta}_0|$ for all outcomes, which naturally follows from our definition of cohorts: the estimated coefficients $\hat{\nu}_0$ capture the average treatment effect in the calendar *week* of coaching, whereas the $\hat{\delta}_0$'s capture the stronger effect on the day of coaching. Second and more importantly, Wald tests show that for all four outcomes the estimated lead coefficients $\nu_{-10}, \ldots, \nu_{-1}$ jointly remain far from significant. This indicates that the parallel pretrends assumption is not violated. Third, the pattern of the dynamic treatment effects does not materially change for any of the outcome variables. For fuel economy, acceleration and – to a lesser extent – cornering, we again observe a short term persistence that peters out after four to seven weeks.

Table 7 reports for all outcome variables the time invariant coaching effect that follows from estimating coefficient δ^{post} in the static specification (5). Panel A shows the estimates when we use a standard TWFE estimator and panel B when we use a heterogeneity robust estimator. Next to the standard p-values, we also report p-values that apply a Bonferroni and a Holm correction for multiple hypothesis testing (MHT). These are conservative methods to adjust for the fact that we consider the impact of coaching on four different outcome variables.[31] As in the dynamic case, the estimates do not materially change when we account for possible treatment heterogeneity. Only for acceleration, we identify an effect that persists for the entire post-coaching period. This consistency suggests that the estimated static treatment effects $\hat{\delta}^{post}$ are unbiased estimates of particular weighted average causal effects. Athey and Imbens (2022) prove that this is always the case whenever the treatment adoption date is randomly assigned.

Re-examinations of studies that apply a staggered adoption design in Baker et al. (2022) show that recently developed alternative estimators like the IW estimator often do not replicate the published results. Hence the more important is our result that the two sets of estimates are highly similar.[32]

[30] We used the Stata package *eventstudyinteract*.

[31] The Bonferroni multiplicity-adjusted p-values are obtained by multiplying the unadjusted p-values by the number of hypotheses (12); the Holm multiplicity-adjusted p-values are obtained by ranking the unadjusted p-values from largest to smallest and to multiply each unadjusted p-value with its rank. Our stratified design disallows us to apply the less conservative MHT correction method developed by List et al. (2019) that assumes simple random matching.

[32] Baker et al. (2022) point out that the biases associated with TWFE-estimates become less problematic the

5.2.1 Heterogeneous Treatment Response

Whereas the previous section dealt with possible treatment effect heterogeneity across adoption cohorts, drivers who are coached at the same date and by the same person may also respond differently to in-person coaching. A prevalent finding in the literature on peer effects in educational outcomes (Sacerdote 2011) is that high-ability students benefit most from the presence of high-ability peers (Hoxby and Weingarth 2005, Lavy et al. 2011, 2012, Fruehwirth 2013), although some studies (Burke and Sass 2013) find that students with the lowest past performance gain most from exposure to higher-achieving peers.[33] In our design, a coaching session explicitly exposes a driver to a high-achieving peer. While recognizing the differences between a school and a work environment – in the nature of the interactions and in the outcomes of interest – the cited studies on peer effects suggest that the response to in-person coaching may be heterogeneous and depend on a driver's own past performance.

To examine whether the results on treatment effect heterogeneity carry over to non-educational contexts and whether coaching helps to decrease the worker-quality gap ($\gamma < 0$, Result 2), we take the following approach. We use the estimated pre-coaching productivity metrics $\hat{\mu}^{EB}$ to group drivers into ex ante low, medium, and high productivity workers. Per outcome, drivers with coaching percentile rank up to 25% ($p(\hat{\mu}^{EB}) \leq 25$) are classified high productivity, those with percentile rank $25 < p(\hat{\mu}^{EB}) < 75$ medium productivity, and those with percentile rank $p(\hat{\mu}^{EB}) \geq 75$ low productivity. Drivers classified as, say, low productivity on acceleration, may be in the medium or high productivity group for braking, concerning, or fuel economy. We separately estimate specification (6) with 'week 0' ($D^0_{i,t} = 1$) the day of coaching and with $D^\tau_{i,t}$ ($\tau = -3, -2, -1, 1, 2, 3$) each denoting a 21-day period (instead of seven) before or after coaching.

Table 8 reports the results. For all outcomes (except braking for which estimated effects are again small and mostly insignificant) coaching generates the largest absolute treatment effect among the group of drivers with the lowest pre-coaching productivity, both at the day of coaching as in subsequent weeks. Take fuel economy as an example. Low productivity drivers attain a fuel reduction of 0.93 liters/100km at the day of coaching, which amounts to closing 39% of the pre-existing gap between low and high productivity workers.[34] For drivers who already

larger the percentage of never-treated units. The sizeable percentage (27%) of never-treated drivers in our sample contributes to explaining the replication success, next to the quasi-random assignment of treatment adoption.

[33] Booij et al. (2017) find that low-ability students benefit from having high-ability peers but that high-ability students are unaffected by their peer group composition.

[34] $\frac{-0.928}{25.287-22.893} = -0.388$. Similarly, for acceleration and cornering, the low-productivity drivers' day-of-coaching

belong to the upper 25% of drivers we hardly find any effect of coaching, independently of which productivity outcome we consider.

Hence, coaching does decrease the worker-quality gap but this effect does not last. The fact that coaching benefits lower performers in particular can be considered a second indication that information transmission is the driving mechanism. We can however not exclude the possibility that lower performers respond more because they view the coaching program as affecting job security. The direction of our result is in contrast to the empirical literature on peer effects in education, which predominantly finds that high-achieving students benefit most from the presence of high-achieving peers. Possible explanations are that high-achieving workers in our setting have little room left for further improvement or are less open to a colleague's feedback.

5.2.2 Peer effects and increased effort norm

Work by among others Silver (2021) suggests that the measured positive performance effects at the day of coaching may not (solely) be caused by information transmission or a signalling effect but result from the fact that drivers experience a higher effort norm when they work with a high achieving peer such as a coach. Our conceptual framework suggests a testable implication of such peer effects (Result 3): when the coaching has an effect through an increase in the salience of the effort norm, one expects a driver to exert more effort in the post-coaching shifts that overlap with the shifts of the coach. We test this by determining, based on the start and end time of their shifts, for all drivers which of their trips have overlap with the shifts of their coaches.[35] In total, 17.3% of a driver's and coach's shifts overlap.

We re-estimate equation (5), interacting the post-coaching indicator functions with a dummy variable for whether or not the driver's trip overlaps with the shift of the coach. Panel A of Table 9 reports per outcome the estimated coefficients together with p-values of an equality of coefficients test. None of these differences is significant such that this exercise does not provide evidence that the effect of coaching operates through a match effect that increases the salience of the effort norm. However, we noted earlier that bus driving is by nature a rather isolated job, which means that even when they have an overlap in shift, workers may in practice not closely collaborate at all. In this, our context differs from the teams of physicians in Emergency

effect amounts to bridging 29.9% and 47.8% of the pre-existing gap, respectively.

[35]For drivers with the role of coach we only observe the start and end time of their shift at the days at which they drive themselves. For days when they act as coach we do not observe this and we assume that they are at the location the entire day.

Departments studied in Silver (2021). There, collaboration is likely to be more direct and intense.

Finally, we examine whether fatigue in post-coaching trips makes it more difficult for drivers to implement the instructions received in a coaching session. As a measure of fatigue, we consider the amount of time a driver has already worked at the start of a new trip. When this is below the median number of 2.82 hours, we classify the worker condition as one of no-fatigue, otherwise we consider the driver fatigued. Panel B of Table 9 shows no differential treatment effect of coaching conditional on a driver's fatigue. Only for the two outcomes without significant post-coaching treatment effects, braking and cornering, a Wald-test rejects the null hypothesis of identical treatment coefficients.

5.2.3 Robustness check: Heterogeneity in Coach Quality

Coaching is provided by a small number of six coaches. One potential worry is that the observed average treatment effects are not caused by an inherent feature of in-person coaching that is independent of who provides the coaching, but is instead due to one or two coaches with idiosyncratic coaching qualities that are hard to copy. In that case, we cannot draw the general conclusion that in-person coaching improves worker productivity. We cannot directly compare differences in how our six coaches provided feedback to drivers because we lack this information. We can however estimate the treatment effect of each individual coach by considering the subsample of drivers instructed by that coach. Substantial heterogeneity in the quality of instructions given by the coaches should result in between-coach differences in treatment effects. Online Appendix H shows the coach-level treatment effects of in-person coaching on fuel economy and ABC outcomes. Despite the fact that the estimates are less precise due to the smaller subsamples, the pattern is remarkably consistent across coaches: the point estimate of fuel savings in liters/100km for example is for all coaches within the range [-0.66, -0.31] on the day of coaching. The diminishing and eventually vanishing of this effect in the four to nine weeks following coaching is also common to all coaches. Based on this, we conclude that the observed effect can be attributed to features inherent to in-person coaching.

5.2.4 Robustness check: Replication In-Person Coaching in Control Region

Over time, the company introduced the EcoManager program in its other concession areas as well. Importantly, the experimental peer-comparison feedback treatments were not part of the implementation in these other areas. We exploit this difference to analyze whether our results on the effects of coaching replicate in an environment where interference between the in-person coaching program and the peer-comparison messages is ruled out by design. We consider the control region South Holland at 200km distance from the treatment region in Friesland. The general written feedback and in-person coaching were introduced in this region on December 1, 2016 (see Figure 1). The data we have for drivers in this region are similar to that for the treatment region and covers over 1.5 mln. trip-level observations by 1,032 tenured drivers in the period 01 July 2015 - 31 Jan. 2017, supplemented with coach logs. We run an analysis similar to the one for the treatment region to evaluate the short- and long-run effects of coaching. Tables 10 and 11 present the results.[36] In sign and magnitude, the estimates are reassuringly similar to those reported in Tables 6 and 7. Fuel economy for example improves by 0.46 liters/100km on the day of coaching. This positive effect persists in the first week following coaching but vanishes in the longer run. For acceleration, the magnitude and dynamic development of the treatment effect is very comparable to the pattern observed in Friesland; for braking and cornering, the treatment effects are a bit more pronounced than in the treatment region.

5.2.5 Final check: Negative Effects on Punctuality

When coaching increases the attention drivers pay to fuel efficient and comfortable driving, this may be detrimental to other elements of the service that a driver is supposed to deliver. One important such other element is punctuality, which we define as the difference between the actual and scheduled trip length. We examine whether coaching negatively impacts a driver's punctuality by estimating equation (6) with punctuality as dependent variable. Regression results show no treatment effect on punctuality, except for a spike at the day of coaching itself.[37] This likely just reflects that the driver and coach indeed interact and discuss the trip.

6 Implications and conclusions

6.1 Implications for research

Our findings contribute to the literatures on peer learning in the workplace and incentive design for energy conservation efforts among workers. First, an unsettled debate in the literature on

[36] Appendix Figure J.11 replicates Figure 3 in plotting for each outcome the dynamic treatment effects.
[37] Appendix Figure L.14 plots the dynamic treatment effects of coaching on punctuality.

peer effects in the workplace is whether productivity improvements are due to social pressure or knowledge spillovers (Cornelissen et al. 2017). Some studies in the lab (Falk and Ichino 2006) and in the field (Mas and Moretti 2009) argue that social pressure is the dominant channel through which peer effects operate, while other research points to knowledge and skills transfers between peers (Papay et al. 2020, Sandvik et al. 2020, Lindquist et al. 2017, Chan et al. 2014). Identifying causality is challenging, as it requires information on the direction of possible spillovers, the time of the interaction, as well as granular data to estimate the impact on individual productivity. Our setting addresses these issues because bus driving is a task in which performance is directly linked to individual effort and because, by design, the direction of the knowledge spillover and when the peer interaction took place is exactly known. We find that peer coaching leads to significant, albeit temporary, improvements in productivity for drivers who initially performed poorly. The evidence is consistent with knowledge and skills transfers between a worker and a high-achieving peer and less with peer pressure through an increased salience of effort norm. The declining effect suggests that the transferred knowledge and skills are subject to decay.

Second, we contribute to the growing literature on designing incentives in the workplace to engage workers in energy conservation efforts. Existing research thus far has mainly focused on residential energy consumption (Gerarden et al. 2017) and attention for conservation efforts on the work floor has been limited (Gosnell et al. 2020). Our results show that stimulating worker-level conservation efforts can be a promising avenue for reducing energy consumption by firms, but that close attention should be paid to the design of the tailored feedback.

6.2 Implications for practice

The managerial implications for the implementation of tailored feedback programs in organizations are as follows. The results suggest that managers can improve worker productivity by having workers and high-achieving peers interact. Our results show that peer coaching has a strong and immediate positive impact on driving behavior, especially among drivers with the lowest initial performance. Although the effect is transient, the decay is not immediate which suggests that productivity improvements are not merely due to peer pressure. Our preferred explanation is that drivers, at least temporarily, perform better because of knowledge and skills transfers between peers. In the weeks following coaching, 17.52 liter of fuel is saved per coached driver, which amounts to €19.27, or €60 per day of coaching. This is less than the cost to free up an experienced driver to coach, but the benefits may outweigh the cost if the company or passengers sufficiently value the improved comfort, environmental gains and personnel development.

To examine the management views within the company, we invited 105 employees at different levels of management to an ex-post survey. The company has defined five regions for its Dutch bus activities, each of which has one regional director and one Eco-coordinator. We invited all these directors and coordinators (ten in total) plus all 46 lower-level team managers and the 49 eco-coaches. We asked them to share their expectations about the in-person coaching program. A high 73% response rate and completed surveys from all concession areas give confidence that the answers are representative.[38] Respondents state that a broad range of benefits should be considered when evaluating the effects of coaching: 83% believe that coaching should be implemented even if the fuel savings are smaller than the cost of the program, but only if coaching leads to improvements on other dimensions as well, such as passenger comfort and personnel development. Hence, to understand why firms implement and keep certain practices, researchers should avoid a narrow focus on pecuniary benefits and costs.

Both coaches and team managers overestimate the fuel savings from a coaching session. They respectively believe that drivers will drive 197 and 250 meters extra per liter of fuel following a coaching session while the actual number is about 116 extra meters. About half (44%) of the respondents expects the effects of coaching to last, whereas the remaining respondents expect the impact to have halved after about six weeks. Our study proves these expectations to be too optimistic: the effect of coaching does not last and has halved after a few weeks. Surveyed employees believe that especially average drivers benefit from coaching. Our empirical findings however show that the positive effect is concentrated in the bottom half of the performance distribution.

These disparities illustrate the importance of using quantitative analysis to check whether management beliefs are correct. Monitoring technologies can assist managers to select workers who would benefit from coaching and to signal when the decay of the coaching effect sets in. Biased beliefs may create inefficiencies in the allocation of coaching capacity when managers have a steering role in the allocation of coaches to drivers. Mentoring programs which more structurally match workers with high-achieving peers may create lasting effects if workers internalize new knowledge and skills and habituate to the improved driving style.

[38]Figures and tables summarizing the key findings of the survey are in online appendix K.

6.3 Concluding Remarks

Our findings give a number of directions for future research. The increased adoption of digital monitoring technologies in workplace settings creates many new opportunities for measuring worker performance and for tailoring feedback to increase their motivation. We have utilized technologies and blended them with recent econometric identification techniques to examine the response to peer coaching. We document clear and heterogeneous responses and believe that more research should be done to investigate how and when peer interactions lead to increased productivity through transfers of knowledge and skills. Designing and evaluating other data-driven incentives could yield fruitful research as we reckon that we have only scratched the surface. Finally, the general question how conservation efforts can be stimulated when someone else pays the bill is in need of more answers.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *Quarterly Journal of Economics*, February 2023, *138* (1), 1–35.

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

____ **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.

Arrow, Kenneth J., "Methodological Individualism and Social Knowledge," *American Economic Review*, 1994, *84* (2), 1–9.

Athey, Susan and Guido W. Imbens, "Design-based analysis in Difference-In-Differences settings with staggered adoption," *Journal of Econometrics*, January 2022, *226* (1), 62–79.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang, "Superstar Extinction," *Quarterly Journal of Economics*, 2010, *125* (2), 549–589.

Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144*, 370–395.

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.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social Incentives in the Workplace," *Review of Economic Studies*, 2010, *77* (2), 417–458.

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," *Review of Economic Studies*, March 2020, *87* (2), 721–749.

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

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek, "Ability Peer Effects in University: Evidence from a Randomized Experiment," *Review of Economic Studies*, 2017, *84*, 547–578.

Brandon, Alec, John A. List, Robert D. Metcalfe, Michael K. Price, and Florian Rundhammer, "Testing for Crowd Out in Social Nudges: Evidence from a Natural Field Experiment in the Market for Electricity," *Proceedings of the National Academy of Sciences*, 2019, *116*, 5293–5298.

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

Burke, Mary A. and Tim R. Sass, "Classroom Peer Effects and Student Achievement," *Journal of Labor Economics*, 2013, *31* (1), 51–82.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225*, 200–230.

Cameron, A. Colin and Douglas L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, Spring 2015, *50* (2), 317–372.

Chan, Tat Y., Jia Li, and Lamar Pierce, "Learning from Peers: Knowledge Transfer and Sales Force Productivity Growth," *Marketing Science*, 2014, *33* (4), 463–484.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson, "Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector," *American Economic Review*, August 2016, *106* (8), 2110–2144.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, September 2014, *104* (9), 2593–2632.

Choudhary, Vivek, Masha Shunko, Serguei Netessine, and Seongjoon Koo, "Nudging Drivers to Safety: Evidence from a Field Experiment," *Management Science*, June 2022, *68* (6), 4196–4214.

Cornelissen, Thomas, Christian Dustmann, and Uta Schoenberg, "Peer Effects in the Workplace," *American Economic Review*, feb 2017, *107* (2), 425–456.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Estimators with Heterogeneous Treatment Effects," *American Economic Review*, September 2020, *110* (110), 2964–2996.

____ and ____, "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey," NBER Working Paper no. 29691 January 2022.

De Grip, Andries and Jan Sauermann, "The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment," *The Economic Journal*, May 2012, *122* (560), 376–399.

Falk, Armin and Andrea Ichino, "Clean Evidence on Peer Effects," *Journal of Labor Economics*, jan 2006, *24* (1), 39–57.

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.

Fruehwirth, Jane Cooley, "Identifying peer achievement spillovers: Implications for desegregation and the achievement gap," *Quantitative Economics*, 2013, *4*, 85–124.

Gerarden, Todd D., Richard G. Newell, and Robert N. Stavins, "Assessing the Energy-Efficiency Gap," *Journal of Economic Literature*, December 2017, *55* (4), 1486–1525.

Gneezy, U., M. Niederle, and A. Rustichini, "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, 2003, *118*, 1049–1074.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225*, 254–277.

Gosnell, Greer K., John. A List, and Robert Metcalfe, "The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains," *Journal of Political Economy*, April 2020, *128* (4), 1195–1233.

Herbst, D. and A. Mas, "Peer Effects on Worker Output in the Laboratory Generalize to the Field," *Science*, 2015, *350*, 545–549.

Holliday, Scott, Jacob LaRiviere, David Novgorodsky, and Michael K. Price, "Prices versus Nudges: What Matters for Search versus Purchase of Energy Investments," *Journal of Public Economics*, 2019, *172*, 151–173.

Hölmstrom, Bengt, "Moral Hazard and Observability," *The Bell Journal of Economics*, 1979, *10* (1), 74–91.

Hoxby, Caroline M. and Gretchen Weingarth, "Taking race out of the equation: School reassignment and the structure of peer effects," `https://www.pausd.org/sites/default/files/pdf-faqs/attachments/TakingRaceOutOfTheEquation.pdf` 2005.

Hubbard, Thomas N., "Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking," *American Economic Review*, 2003, *93* (4), 1328–1353.

ICCT, "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 Elias Bruegmann, "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics*, oct 2009, *1* (4), 85–108.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan, "Self-Control at Work," *Journal of Political Economy*, dec 2015, *123* (6), 1227–1277.

Kelley, Erin M., Gregory Lane, and David Schönholzer, "The Impact of Monitoring Technologies on Contracts and Employee Behavior: Experimental Evidence from Kenya's Transit Industry," mimeo, University of Berkeley 2018.

Larrick, Richard P. and Jack B. Soll, "The MPG Illusion," *Science*, 2008, *320* (5883), 1593–1594.

Lavy, Victor, M. Daniele Paserman, and Analia Schlosser, "Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom," *The Economic Journal*, March 2011, *122*, 208–237.

——, **Olmo Silva, and Felix Weinhardt**, "The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools," *Journal of Labor Economics*, 2012, *30* (2), 367–414.

Lazear, Edward P., Kathryn L. Shaw, and Christopher Stanton, "Making Do with Less: Working Harder during Recessions," *Journal of Labor Economics*, 2016, *34* (1), S333–S360.

Lindquist, Matthew, Jan Sauermann, and Yves Zenou, "Peer Effects in the Workplace: A Network Approach," *CEPR Discussion Paper 10928*, 2017.

List, Johan A., Azeem M. Shaikh, and Yang Xu, "Multiple hypothesis testing in experimental economics," *Experimental Economics*, 2019, *22*, 773–793.

Mas, Alexandre and Enrico Moretti, "Peers at Work," *American Economic Review*, mar 2009, *99* (1), 112–145.

Morris, Carl N., "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, March 1983, *78* (381), 47–55.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary E. Laski, "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data," *American Economic Journal: Economic Policy*, feb 2020, *12* (1), 359–388.

Pierce, Lamar, Daniel C. Snow, and Andrew McAfee, "Cleaning House: The Impact of Information Technology Monitoring on Employee Theft and Productivity," *Management Science*, October 2015, *61* (10), 2299–2319.

Romensen, Gert-Jan and Adriaan R. Soetevent, "Does Targeting Relative Performance Feedback Improve Worker Productivity? Field Experimental Evidence from Bus Drivers," University of Groningen working paper (link) 2024.

Sacerdote, Bruce, "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?," in Ludger Woessmann Eric A. Hanushek, Stephen Machin, ed., *Handbook of the Economics of Education*, Vol. 3, Elsevier, 2011, chapter 4, pp. 249–277.

Sandvik, Jason, Richard Saouma, Nathan Seegert, and Christopher Stanton, "Workplace Knowledge Flows," *The Quarterly Journal of Economics*, August 2020, *135* (3), 1635–1680.

Silver, David, "Haste or Waste? Peer Pressure and Productivity in the Emergency Department," *Review of Economic Studies*, 2021, *88*, 1385–1417.

Soleymanian, Miremad, Charles B. Weinberg, and Ting Zhu, "Sensor Data and Behavioral Tracking: Does Usage-Based Auto Insurance Benefit Drivers?," *Marketing Science*, 2019, *38* (1), 21–43.

Staats, Bradley R., Hengchen Dai, David Hofmann, and Katherine L. Milkman, "Motivating Process Compliance Through Individual Electronic Monitoring: An Empirical Examination of Hand Hygiene in Healthcare," *Management Science*, May 2017, *63* (5), 1563–1585.

Sun, Liyang and Jesse M. Shapiro, "A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure," *Journal of Economic Perspectives*, Fall 2022, *36* (4), 193–204.

—— **and Sarah Abraham**, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225*, 175–199.

Varian, Hal R., *Microeconomic Analysis*, 3rd ed., New York, NY: W. W. Norton & Company, 1992.

Vivalt, Eva, "How much can we generalize from impact evaluations?," *Journal of the European Economic Association*, 2020, *18* (6), 3045–3089.

Waldinger, Fabian, "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany," *Review of Economic Studies*, 2012, *79* (2), 838–861.

Wooldridge, Jeffrey M., "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," Available at SSRN: `http://dx.doi.org/10.2139/ssrn.3906345` August 2021.

Tables and Figures

Tables

Table 1: Quasi-Random Coaching: Balance Tests

	C	NC	Δ(C-NC)	stand. p-value	Bonf. corr.	Holm corr.
Baseline performance						
Fuel economy	25.038	25.247	-0.209	0.0865	1.0000	0.2595
Acceleration	13.132	13.349	-0.217	0.5200	1.0000	1.0000
Braking	3.715	3.789	-0.075	0.7744	1.0000	1.0000
Cornering	1.093	1.189	-0.096	0.2065	1.0000	1.0000
Demographics						
Year of birth	1962.4	1961.8	0.558	0.3372	1.0000	1.0000
Year of employment	1996.2	1996.3	-0.150	0.8706	1.0000	1.0000
Share of FTE≥0.9	0.801	0.776	0.025	0.6205	1.0000	1.0000
Share of female drivers	0.094	0.073	0.020	0.554	1.0000	1.0000
Trip-specific variables						
Punctuality	-2.942	-3.018	0.076	0.2251	1.0000	1.0000
Distance traveled (in km.)	30.564	32.025	-1.461	0.0996*	1.0000	0.3984
Number of passengers	15.035	15.283	-0.248	0.5983	1.0000	1.0000
Number of bus stops	37.934	37.864	0.070	0.9268	1.0000	1.0000
Share of rides:						
- Morning rush hours	0.298	0.159	0.139	0.0076***	0.1976	0.0076***
- Evening rush hours	0.125	0.253	-0.127	0.0084***	0.2184	0.0168**
- Weekend	0.031	0.031	0.000	1.0000	1.0000	1.0000
- Fill in	0.004	0.025	-0.021	0.1539	1.0000	0.7695
- Holidays	0.119	0.125	-0.006	0.8839	1.0000	1.0000
- Urban area	0.149	0.146	0.003	0.9366	1.0000	1.0000
- School	0.004	0.004	0.000	0.9954	1.0000	1.0000
Share of rides on bus types						
VDL	0.773	0.755	0.017	0.7427	1.0000	1.0000
Intouro	0.078	0.099	-0.021	0.5555	1.0000	1.0000
IRIS	0.149	0.146	0.003	0.9366	1.0000	1.0000

Notes: Unit of observation is the driver. For every coaching date, the mean baseline performance and non-performance related variables of drivers who receive their first coaching (C) is compared to that of non-yet-coached colleagues (NC). The baseline period spans the pre-announcement period. Reported are the mean values over all coaching dates. Standard p-values as well as p-values that use a Bonferroni and Holm correction for multiple hypothesis testing. Fuel economy in liters/100km; Performance on the ABC dimensions as the number of events/10km. Fewer events mean better driving behavior. Punctuality is the difference in minutes between actual and planned driving time. Number of passengers is based on check-ins with public transport cards. Morning and evening rush hours are from 7-10am and 4-7pm, respectively. Holiday rides take place during public and school holidays. School rides are along routes with schools and universities as final destinations. Fill-ins are non-scheduled trips whereby a driver replaces a colleague from another base location.
***(**,*) : the corresponding p-values are less than 1% (5% or 10%).

Table 2: Summary Statistics Variation in Productivity Metrics across Drivers

Productivity metric:	Acceleration	Braking	Cornering	Fuel economy
	(1)	(2)	(3)	(4)
	Panel A. Driving condition adjusted quality measures			
Average value	12.320	2.278	2.442	24.379
	(2.392)	(1.798)	(2.101)	(1.200)
Dispersion $\Delta(p_{90} - p_{10})$	4.630	1.252	2.415	2.668
Coeff. of variation (c_v)	0.1940	0.7890	0.8600	0.0490
Attenuation factor (average)	0.015	0.034	0.032	0.005
Attenuation factor (median)	0.002	0.001	0.000	0.001
No. trips per driver	317.34	317.34	317.34	690.13
No. drivers	372	372	372	405
	Panel B. Correlation of quality metrics			
Productivity metric:	Acceleration	Braking	Cornering	Fuel economy
Acceleration	1.00			
	[372]			
Braking	0.34	1.00		
	[372]	[372]		
Cornering	0.34	0.58	1.00	
	[372]	[372]	[372]	
Fuel economy	0.64	0.39	0.30	1.00
	[371]	[371]	[371]	[405]
	Panel C. Rank of coaches			
Productivity metric:	Acceleration	Braking	Cornering	Fuel economy
Minimum	0.27	0.27	0.27	1.73
Median	6.05	10.62	19.89	8.15
Maximum	15.05	35.75	35.22	24.94

Notes: The sample considers the time period with complete coach logs (01/01/2015-30/04/2016) and includes all pre-coaching observations from drivers with at least 25 pre-coaching trip-level observations. Summary statistics are reported across drivers. Fuel economy: liters/100km. ABC dimensions: no. of events/10km. Panel A: The standard deviations are of the driving condition adjusted measures and are EB-adjusted to account for measurement error (see online Appendix B for details). $\Delta(p_{90} - p_{10})$: difference in EB-adjusted quality measure 90th vs. 10th percentile driver. c_v: coefficient of variation. Panel B: The number in brackets denotes the number of drivers used to compute correlation. Panel C: Per outcome, the percentile ranks based on the $\hat{\mu}_i^{EB}$'s are reported.

Table 3: Effect of Announcement EcoManager Campaign on Fuel Economy

Period	Control region	Treatment region	Difference
Baseline (01/07/15-08/11/15)	27.677	24.155	-3.522
	(0.080)	(0.076)	(0.062)
	[225,660]	[96,412]	[322,072]
Post-announcement (09/11/15-14/12/15)	27.821	24.187	-3.634
	(0.105)	(0.088)	(0.087)
	[62,436]	[25,475]	[87,911]
Difference over time	0.144	0.032	DID_{Ann} = -0.112
	(0.131)	(0.116)	(0.107)
	[288,096]	[121,887]	[409,983]

Notes: Each cell shows for various subsets of the sample the mean number of liters of fuel used to drive 100km per driver and per trip. Standard errors (clustered by date) in parentheses, the number of observations in square brackets. The baseline period spans 01/07/15 to 08/11/15. The post-announcement period spans 09/11/15 to 14/12/15.

Table 4: Effect of Introduction General Feedback on Fuel Economy

Period	Control region	Treatment region	Difference
Baseline (01/07/15-14/12/15)	27.708	24.161	-3.547
	(0.067)	(0.063)	(0.052)
	[288,096]	[121,887]	[409,983]
Post-feedback (15/12/15-30/04/16)	27.280	23.661	-3.619
	(0.061)	(0.047)	(0.044)
	[238,256]	[118,514]	[356,770]
Difference over time	-0.428	-0.500	DID_{GenFB} = -0.072
	(0.091)	(0.078)	(0.076)
	[526,352]	[240,401]	[766,753]

Notes: Each cell shows for various subsets of the sample the mean number of liters of fuel used to drive 100km per driver and per trip. Standard errors (clustered by date) in parentheses, the number of observations in square brackets. The baseline period spans 01/07/15 to 14/12/15. The post-feedback period spans 15/12/15 to 30/04/16.

Table 5: Effect Introduction General Feedback on Fuel Economy and Comfort

	(1)	(2)	(3)	(4)
		A. Fuel Economy		
$\hat{\gamma}$ (PF)	-0.490***	-0.282***	0.002	-0.109
	(0.068)	(0.073)	(0.077)	(0.080)
Constant	27.558***	23.749***	26.690***	26.733***
	(0.080)	(0.057)	(0.131)	(0.126)
R^2	.101	.432	.439	.447
Number of trip-level observations	734242	734242	734242	734242
Controls	No	Yes	Yes	Yes
Driver fixed effects	No	Yes	Yes	Yes
Day fixed effects	No	No	Yes	Yes
Bus type × day fixed effects	No	No	No	Yes
		B. Comfort		
Dependent variable:		**Acceleration**	**Braking**	**Cornering**
$\hat{\gamma}$ (PF)		-2.226***	-0.869***	-0.408***
		(0.151)	(0.038)	(0.052)
Constant		9.942***	1.307***	1.073***
		(0.216)	(0.049)	(0.090)
R^2		.067	.151	.032
Number of trip-level observations		189626	189626	189626
Controls		No	No	No
Driver fixed effects		No	No	No
Day fixed effects		No	No	No
Bus type × day fixed effects		No	No	No

Notes: Identification of the post-feedback effect on driving performance. The time period under consideration is from 01/01/2015 to 30/04/2016. Fuel economy is measured in liters/100km. The ABC outcome variables are measured as the number of events per 10 kilometers. Standard errors are clustered by driver which is the unit of randomization (Abadie et al. 2023). Controls include: travel distance, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, non-scheduled rides, having been coached and treatment differences in written feedback. A no-report indicator is included to capture drivers operating after December 15, 2015, but who have not yet received their first report. ***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

Table 6: Dynamic Treatment Effects In-Person Coaching on Driving Performance

Dep. var.	Fuel Economy		Acceleration		Braking		Cornering	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	TWFE	SA-IW	TWFE	SA-IW	TWFE	SA-IW	TWFE	SA-IW
	$\hat{\delta}_\tau$	$\hat{\nu}_\ell$	$\hat{\delta}_\tau$	$\hat{\nu}_\ell$	$\hat{\delta}_\tau$	$\hat{\nu}_\ell$	$\hat{\delta}_\tau$	$\hat{\nu}_\ell$
$\tau, \ell = -10$	0.074	-0.101	-0.091	0.124	-0.009	-0.003	-0.072	-0.075
	(0.083)	(0.105)	(0.160)	(0.222)	(0.043)	(0.047)	(0.048)	(0.059)
-9	0.111	0.112	0.086	0.204	0.046	-0.007	0.000	-0.011
	(0.074)	(0.097)	(0.153)	(0.173)	(0.046)	(0.048)	(0.048)	(0.056)
-8	0.089	0.129	0.084	0.358*	0.038	-0.003	-0.037	-0.001
	(0.074)	(0.101)	(0.154)	(0.189)	(0.047)	(0.054)	(0.049)	(0.055)
-7	0.032	0.028	-0.058	0.121	0.011	-0.004	0.016	-0.002
	(0.070)	(0.101)	(0.130)	(0.165)	(0.039)	(0.048)	(0.048)	(0.051)
-6	0.071	0.012	-0.032	0.124	0.032	0.039	-0.015	-0.023
	(0.068)	(0.091)	(0.128)	(0.171)	(0.041)	(0.043)	(0.042)	(0.045)
-5	0.029	0.108	-0.062	0.125	0.072*	0.05	0.005	0.036
	(0.065)	(0.092)	(0.119)	(0.15)	(0.038)	(0.041)	(0.038)	(0.042)
-4	0.021	0.014	0.034	0.273*	0.018	0.037	0.066*	0.039
	(0.060)	(0.078)	(0.126)	(0.144)	(0.037)	(0.035)	(0.034)	(0.036)
-3	0.035	-0.033	-0.158	0.057	0.058*	0.035	0.040	0.074*
	(0.068)	(0.089)	(0.123)	(0.151)	(0.031)	(0.037)	(0.037)	(0.041)
-2	0.039	0.032	-0.184	0.027	0.035	0.046	-0.007	0.029
	(0.061)	(0.081)	(0.119)	(0.149)	(0.030)	(0.037)	(0.027)	(0.032)
-1	0	0	0	0	0	0	0	0
	()	()	()	()	()	()	()	()
0	-0.507***	-0.396***	-1.060***	-0.508***	-0.078**	-0.019	-0.188***	-0.064**
	(0.093)	(0.087)	(0.167)	(0.137)	(0.033)	(0.024)	(0.035)	(0.028)
1	-0.230***	-0.258***	-0.699***	-0.446***	-0.002	0.011	-0.102***	0.000
	(0.063)	(0.09)	(0.122)	(0.155)	(0.025)	(0.026)	(0.028)	(0.031)
2	-0.100	-0.214**	-0.507***	-0.263	-0.033	-0.003	-0.025	-0.02
	(0.069)	(0.091)	(0.135)	(0.165)	(0.026)	(0.027)	(0.029)	(0.035)
3	-0.201***	-0.336***	-0.540***	-0.686***	0.007	0.023	-0.077***	-0.091***
	(0.072)	(0.095)	(0.136)	(0.17)	(0.024)	(0.025)	(0.029)	(0.032)
4	-0.172**	-0.344***	-0.697***	-0.426**	-0.000	0.019	-0.096***	-0.053
	(0.077)	(0.106)	(0.144)	(0.173)	(0.025)	(0.028)	(0.031)	(0.036)
5	-0.081	-0.239**	-0.407***	-0.03	0.009	0.006	-0.052	0.008
	(0.077)	(0.113)	(0.135)	(0.195)	(0.026)	(0.028)	(0.032)	(0.037)
6	-0.090	-0.231*	-0.383**	-0.312*	0.026	0.046	-0.021	-0.047
	(0.083)	(0.118)	(0.151)	(0.188)	(0.027)	(0.032)	(0.037)	(0.041)
7	-0.120	-0.273**	-0.405**	-0.207	0.038	0.038	-0.037	0.011
	(0.084)	(0.114)	(0.157)	(0.195)	(0.031)	(0.031)	(0.037)	(0.046)
8	-0.057	0.036	-0.426**	0.043	0.031	0.045	0.012	0.035
	(0.081)	(0.117)	(0.171)	(0.215)	(0.028)	(0.031)	(0.043)	(0.046)
9	0.003	-0.102	-0.397**	-0.166	0.047	0.023	-0.019	0.005
	(0.083)	(0.122)	(0.166)	(0.219)	(0.030)	(0.034)	(0.043)	(0.05)
10	0.002	-0.064	-0.376**	-0.265	0.034	0.018	-0.005	-0.025
	(0.087)	(0.118)	(0.186)	(0.229)	(0.031)	(0.035)	(0.043)	(0.050)
Wald test								
leads $_{\tau,\ell=-10,\dots,-2}$	0.945	0.637	0.502	0.237	0.415	0.504	0.043	0.847
lags $_{\tau,\ell=0,\dots,10}$	<0.0001	0.007	<0.0001	0.052	0.064	0.440	<0.0001	0.503
Obs. (trips)	352,253	352,253	187,127	187,127	187,127	187,127	187,127	187,127
Nr. drivers	399	399	376	376	376	376	376	376
Controls	Yes	No	Yes	No	Yes	No	Yes	No

Notes: This table reports the dynamic treatment effects of in-person coaching on driving performance. The considered time period is the period with complete logs available from all coaches (01/01/2015-30/04/2016).
Columns (1), (3), (5), (7): TWFE estimates. $\hat{\delta}_0$ is the estimated effect at the day of coaching, $\hat{\delta}_1$ ($\hat{\delta}_{-1}$) the estimated effect 1-7 days after (before) coaching etc.
Columns (2), (4), (6), (8): Sun and Abraham (2021)'s $\hat{\nu}_\ell$ IW estimates that are weighted averages of the DID estimators of the cohort-specific average treatment effects $CATT_{e,\ell}$, $e \in E$ with E the set of cohorts. $\hat{\nu}_0$ is the estimated effect in the week of coaching, $\hat{\nu}_1$ ($\hat{\nu}_{-1}$) the estimated effect the first week after (before) coaching etc.
The dependent variable fuel economy is measured in liters/100km, acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver and shown in parentheses. The omitted category is the seven days (week) before coaching. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, and non-scheduled rides.
$^{***}(^{**},^{*})$: statistically different from zero at the 1%-level (5%-level, 10%-level).

Table 7: Time Invariant Effect In-Person Coaching on Driving Performance

Outcome	Post-coaching δ^{post}	p-values Unadj.	Bonf.	Holm
A. Standard TWFE				
Fuel economy	-0.1271	0.0285**	0.1140	0.0656*
Acceleration	-0.4486	0.0003***	0.0012***	0.0003***
Braking	-0.0126	0.6171	1.0000	1.0000
Cornering	-0.0442	0.2480	0.9920	0.7440
B. Heterogeneity Robust SA-IW Estimator				
Fuel economy	-0.1552	0.0140**	0.0560	0.0280**
Acceleration	-0.3179	0.0124**	0.0496*	0.0124**
Braking	-0.0145	0.6206	1.0000	1.0000
Cornering	-0.0266	0.5346	1.0000	1.0000

Notes: Identification of in-person coaching effects on driving performance. The time period considered is the period for which we have complete logs available from all coaches (01/01/2015-30/04/2016). Fuel economy is measured in liters/100km; acceleration, braking and cornering as the number of events per 10 kilometers. Standard p-values and p-values that use a Bonferroni and Holm correction for multiple hypothesis testing are reported. Controls, driver, date and bus type×date fixed are included. Standard errors are clustered by driver. ***(**,*) : the corresponding p-values are less than 1% (5% or 10%).

Table 8: Heterogeneous Treatment Effect Coaching Conditional on Pre-Coaching Productivity

	(1)	(2)	(3)	(4)	(5)	(6)
Pre-coaching productivity:	**Lower 25%**	**Middle 50%**	**Upper 25%**	**Lower 25%**	**Middle 50%**	**Upper 25%**
	A. Acceleration			**B. Braking**		
42 – 22 days before	0.490**	0.050	-0.057	0.129**	0.007	0.020
	(0.208)	(0.098)	(0.154)	(0.059)	(0.033)	(0.035)
21 – 1 day before	0	0	0	0	0	0
	()	()	()	()	()	()
Day of coaching	-1.185***	-1.012***	-0.409	-0.116	-0.143***	-0.026
	(0.378)	(0.211)	(0.329)	(0.115)	(0.045)	(0.036)
1 – 21 days after	-0.665***	-0.555***	-0.080	-0.024	-0.054*	0.017
	(0.252)	(0.136)	(0.187)	(0.056)	(0.028)	(0.028)
22 – 42 days after	-1.094***	-0.351**	0.162	-0.030	-0.058**	0.024
	(0.329)	(0.149)	(0.206)	(0.053)	(0.029)	(0.032)
43 – 63 days after	-1.137***	-0.278	0.363	0.027	-0.025	0.009
	(0.388)	(0.178)	(0.256)	(0.058)	(0.033)	(0.039)
Mean baseline level[1]	10.425	8.694	6.461	1.123	0.743	0.438
21 – 1 day before	(0.398)	(0.203)	(0.338)	(0.113)	(0.060)	(0.036)
No. trip-level obs.	42,977	100,197	42,673	35,447	100,696	49,704
No. drivers	40	86	44	27	81	62
	C. Cornering			**D. Fuel economy**		
42 – 22 days before	0.052	0.030	-0.027*	0.135	-0.028	0.070
	(0.070)	(0.028)	(0.014)	(0.085)	(0.052)	(0.074)
21 – 1 day before	0	0	0	0	0	0
	()	()	()	()	()	()
Day of coaching	-0.590***	-0.191***	-0.026	-0.928***	-0.464***	-0.352**
	(0.104)	(0.036)	(0.031)	(0.253)	(0.109)	(0.153)
1 – 21 days after	-0.208***	-0.095***	-0.013	-0.253**	-0.218***	-0.108
	(0.073)	(0.029)	(0.017)	(0.102)	(0.065)	(0.094)
22 – 42 days after	-0.262***	-0.076**	-0.014	-0.219	-0.175**	0.045
	(0.081)	(0.033)	(0.021)	(0.147)	(0.076)	(0.125)
43 – 63 days after	-0.206*	-0.043	-0.002	-0.046	-0.126	0.036
	(0.110)	(0.043)	(0.026)	(0.181)	(0.085)	(0.124)
Mean baseline level[1]	1.467	0.671	0.232	25.287	24.031	22.893
21 – 1 day before	(0.119)	(0.043)	(0.024)	(0.166)	(0.099)	(0.144)
No. trip-level obs.	39,574	101,717	44,556	77,691	190,880	83,682
No. drivers	34	89	47	30	93	45
Controls		Yes			Yes	
Driver fixed effects		Yes			Yes	
Day fixed effects		Yes			Yes	
Bus type × day fixed effects		Yes			Yes	

Notes: [1]: Mean baseline level: no. of events/10km for Acceleration, Braking, and Cornering; liters/100km for Fuel Economy.
TWFE dynamic treatment estimates of equation (6) (with weeks replaced by three week periods) on subsamples that condition on the EB-adjusted pre-coaching worker productivity estimates $\hat{\mu}^{EB}$ for the outcome considered. Only drivers with observations in all time periods between 21 days before and 42 days after coaching are included. Columns (1) and (4) show the treatment effects for drivers with low pre-coaching productivity (percentile rank $p(\hat{\mu}^{EB}) \geq 75$); columns (2) and (5) idem for workers with medium productivity ($25 < p(\hat{\mu}^{EB}) < 75$); columns (3) and (6) idem for workers with high productivity ($p(\hat{\mu}^{EB}) \leq 25$). The seven days before coaching are the omitted category. "Week 0" is the day coaching is received. Standard errors are clustered by driver. Controls include: travel distance, number of passengers and bus stops, dummies for non-scheduled rides, bus types, morning and evening rush hours, and the interaction of bus type and day fixed effects. Coaches themselves are excluded. Online Appendix Figure I.10 shows coefficient plots of the results. The considered time period: 01/01/2015-30/04/2016.
***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

Table 9: Post-Coaching Heterogenous Treatment Effect Tests

Productivity metric:	Acceleration (1)	Braking (2)	Cornering (3)	Fuel economy (4)
	Panel A. Salience of Effort Norm			
Coach driver do not share shift	-0.434***	-0.010	-0.039	-0.125**
	(0.126)	(0.026)	(0.039)	(0.059)
Coach driver do share shift	-0.471***	-0.017	-0.053	-0.130
	(0.122)	(0.026)	(0.038)	(0.059)
p-values	0.392	0.329	0.168	0.832
	Panel B. Fatigue			
No Fatigue	-0.489***	-0.037	-0.079	-0.148**
	(0.128)	(0.026)	(0.039)	(0.060)
Fatigue	-0.420***	0.005	-0.019	-0.112*
	(0.123)	(0.025)	(0.038)	(0.060)
p-values	0.192	< 0.0001	< 0.0001	0.203
obs.	187,127	187,127	187,127	352,253

Notes: Identification of in-person coaching effects on driving performance. The sample considers the time period with complete coach logs (01/01/2015-30/04/2016). Reported are the estimated δ^{Post} coefficients of equation (5) interacted with dummy variables: Driver and coach share shift (Panel A); Fatigue driver (Panel B), together with the p-values of equality of coefficients Wald tests. Fuel economy is measured in liters/100km; acceleration, braking and cornering as the number of events per 10 kilometers. Controls, driver, date and bus type×date fixed are included. Standard errors are clustered by driver. ***(**,*) : the corresponding p-values are less than 1% (5% or 10%).

Table 10: Dynamic Treatment Effects In-Person Coaching on Driving Performance – Replication Control Region

Dep. var.	**Fuel Economy**	**Acceleration**	**Braking**	**Cornering**
	(1)	(2)	(3)	(4)
	$\hat{\delta}_\tau$	$\hat{\delta}_\tau$	$\hat{\delta}_\tau$	$\hat{\delta}_\tau$
$\tau = -10$	0.028	0.119	0.042	0.069
	(0.134)	(0.157)	(0.059)	(0.058)
-9	-0.181*	0.173	0.018	0.081
	(0.102)	(0.146)	(0.058)	(0.059)
-8	0.010	0.096	0.047	0.111*
	(0.113)	(0.150)	(0.054)	(0.059)
-7	-0.261**	-0.060	-0.001	0.030
	(0.106)	(0.154)	(0.050)	(0.053)
-6	-0.159	-0.016	-0.000	0.041
	(0.118)	(0.169)	(0.051)	(0.054)
-5	-0.168	0.097	0.006	0.045
	(0.104)	(0.154)	(0.048)	(0.057)
-4	-0.146	-0.065	-0.023	0.023
	(0.103)	(0.142)	(0.050)	(0.054)
-3	-0.044	-0.147	0.016	0.016
	(0.108)	(0.139)	(0.048)	(0.049)
-2	-0.157	-0.101	-0.038	-0.036
	(0.100)	(0.131)	(0.042)	(0.042)
-1	0	0	0	0
	()	()	()	()
0	-0.459***	-1.492***	-0.307***	-0.296***
	(0.130)	(0.178)	(0.057)	(0.050)
1	-0.279***	-0.655***	-0.163***	-0.175***
	(0.102)	(0.127)	(0.044)	(0.040)
2	-0.288***	-0.715***	-0.187***	-0.177***
	(0.107)	(0.145)	(0.047)	(0.047)
3	-0.202*	-0.391***	-0.180***	-0.119**
	(0.105)	(0.143)	(0.048)	(0.049)
4	-0.253**	-0.497***	-0.131***	-0.080*
	(0.102)	(0.141)	(0.049)	(0.047)
5	-0.234**	-0.381***	-0.184***	-0.103*
	(0.106)	(0.140)	(0.049)	(0.056)
6	-0.101	-0.366***	-0.125**	-0.097*
	(0.108)	(0.140)	(0.052)	(0.055)
7	-0.069	-0.405***	-0.109*	-0.118**
	(0.131)	(0.157)	(0.057)	(0.051)
8	0.024	-0.350**	-0.144***	-0.075
	(0.128)	(0.163)	(0.053)	(0.053)
9	0.062	-0.374**	-0.139**	-0.090
	(0.121)	(0.162)	(0.056)	(0.064)
10	-0.083	-0.251*	-0.118**	-0.029
	(0.104)	(0.149)	(0.048)	(0.055)
Wald test				
leads $_{\tau=-10,\ldots,-2}$	0.0863	0.6035	0.9131	0.6133
lags $_{\tau=0,\ldots,10}$	0.0001	<0.0001	0.0001	<0.0001
Obs. (trips)	1,788,720	1,100,408	1,100,408	1,100,408
Nr. drivers	1,023	1,025	1,025	1,025
Controls	Yes	Yes	Yes	Yes

This table reports the dynamic treatment effects of in-person coaching on driving performance. TWFE estimates. $\hat{\delta}_0$ is the estimated effect at the day of coaching, $\hat{\delta}_1$ ($\hat{\delta}_{-1}$) the estimated effect 1-7 days after (before) coaching etc. The considered time period is 01/07/15-31/01/17. The dependent variable fuel economy is measured in liters/100km, acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver and shown in parentheses. The omitted category is the seven days (week) before coaching. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, morning and evening rush hours, and non-scheduled rides.
***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

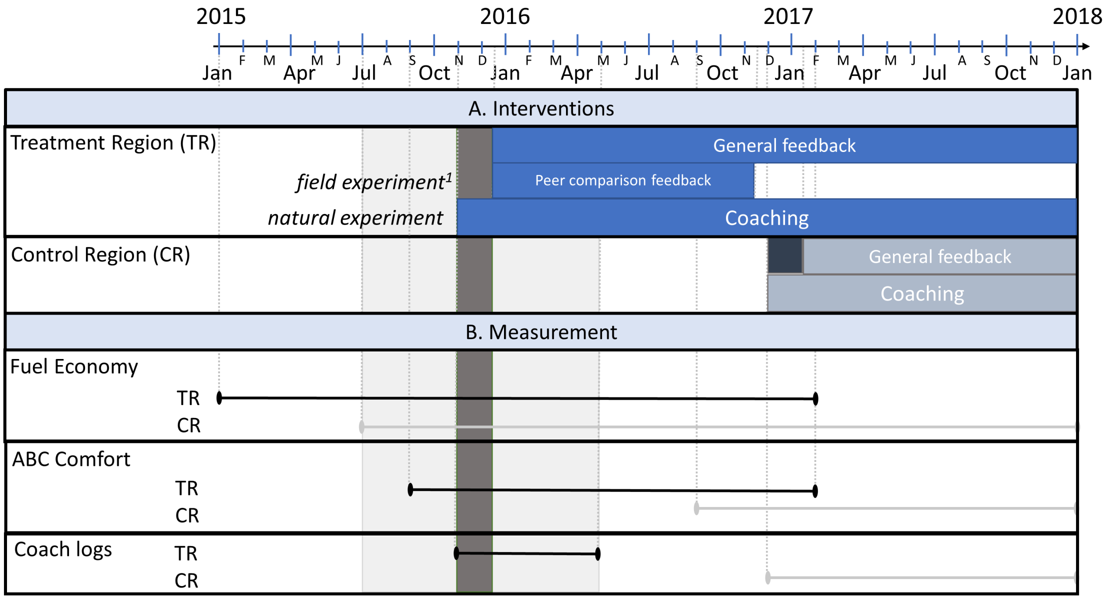
Table 11: Time Invariant Effect In-Person Coaching on Driving Performance – Replication Control Region

Outcome	Post-coaching	p-values		
	δ^{post}	Unadj.	Bonf.	Holm
Fuel economy	-0.0551	0.2542	1.0000	1.0000
Acceleration	-0.4828	0.0000***	0.0000***	0.0000***
Braking	-0.1638	0.0000***	0.0000***	0.0000***
Cornering	-0.1568	0.0009***	0.0036***	0.0027***

Notes: Identification of in-person coaching effects on driving performance in the control region (South Holland). The time period considered is the period from 01/07/2015-31/12/2017. Further: see notes to Table 7. ***(**,*) : the corresponding p-values are less than 1% (5% or 10%).

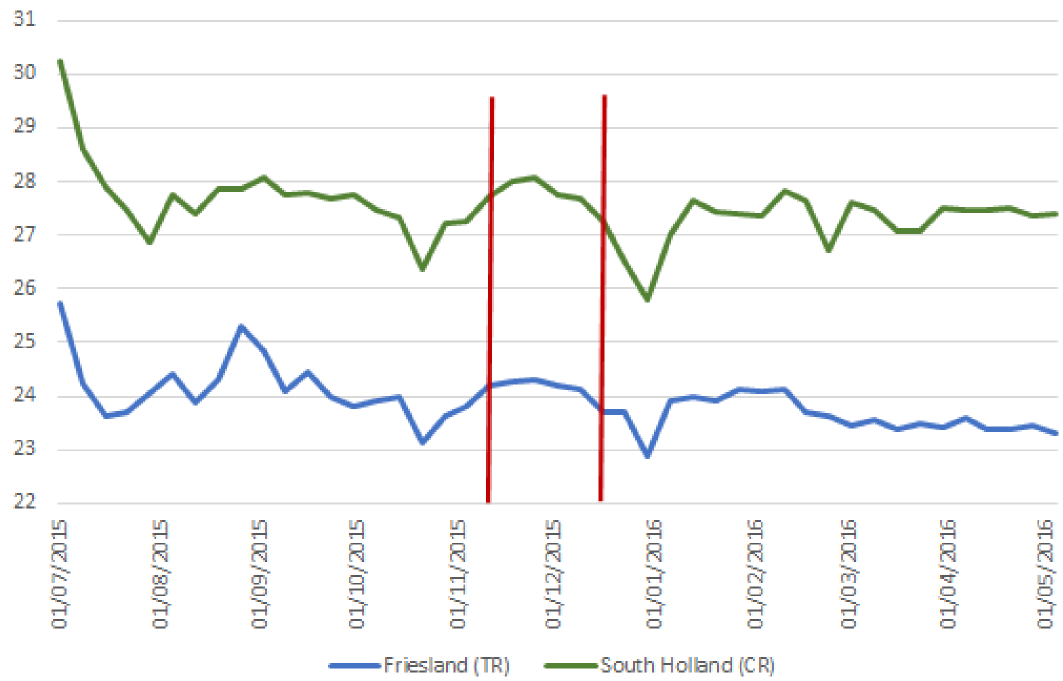
Figures

Figure 1: Timeline Study



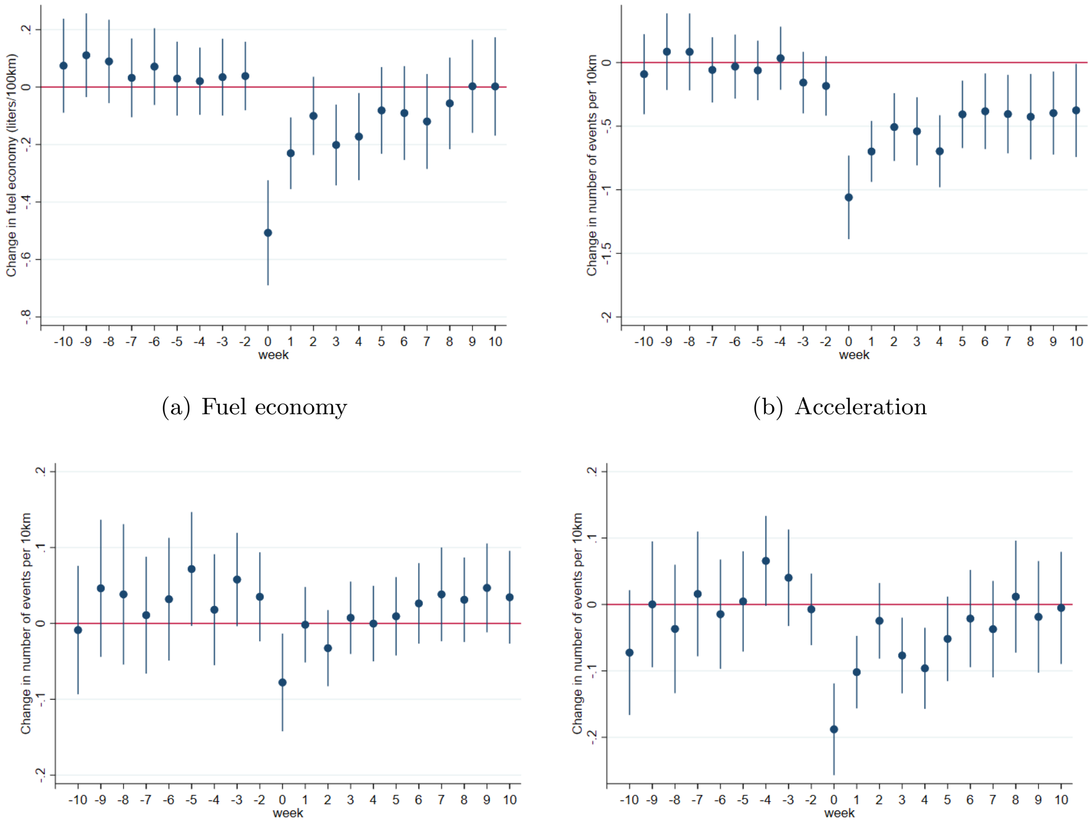
Notes: Panel A shows for the treatment and control region when interventions were announced and implemented. Panel B shows when measurement of the fuel economy and ABC outcome variables and the recording of coach logs started and ended in the different regions. The light-shaded area is the time period for which complete coach logs (after coaching has been introduced in the region) and fuel economy data are available for all drivers in the treatment and control region. The analysis focuses on this period, which runs from 01/07/15 to 30/04/16 and can be divided into a baseline (01/07/15 - 08/11/15), a post-announcement (09/11/15-14/12/15, the dark-shaded area) and a post-feedback period (15/12/15-30/04/16).
[1] : See Romensen and Soetevent (2024) for full details on the design, implementation and results of the field experiment.

Figure 2: Weekly Fuel Economy Averages (liters/100km) in the Treatment and Control Region



Notes: The first vertical red bar shows the EcoManager announcement date in the Treatment Region (09/11/15); the second red bar shows the date that the first feedback reports were distributed in the Treatment Region (15/12/15).

Figure 3: Dynamic Treatment Effects of In-Person Coaching [TWFE]



(a) Fuel economy

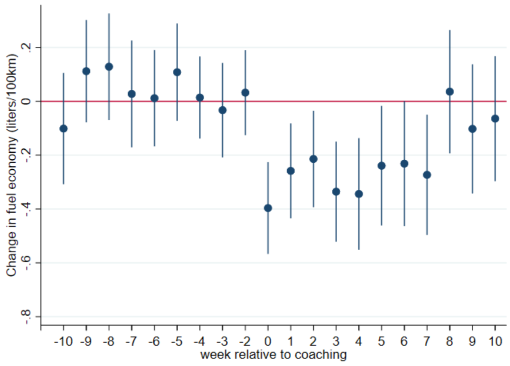
(b) Acceleration

(c) Braking

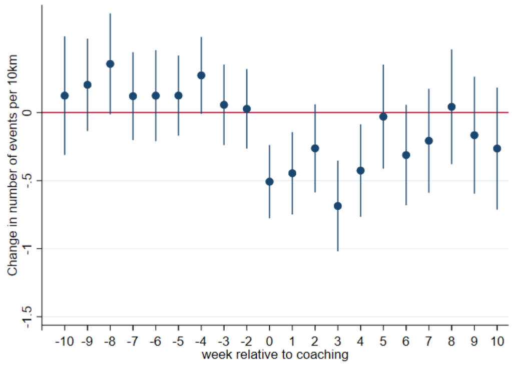
(d) Cornering

Notes: Conventional TWFE dynamic treatment estimates $\hat{\delta}_t$. Driving performance in the 10 weeks before and after coaching based on trips with VDL and Intouro buses. The day of coaching itself is point 0 on the x-axis. The seven days before coaching are the omitted category. The vertical spikes indicate 95% confidence intervals. The dependent variable fuel economy is measured in liters/100km, and acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, number of passengers and bus stops, dummies for non-scheduled rides, bus types, morning and evening rush hours, and the interaction of bus type and day fixed effects. Coaches themselves are excluded from the sample. The considered time period is the period with complete logs available from all coaches (01/01/2015-30/04/2016).

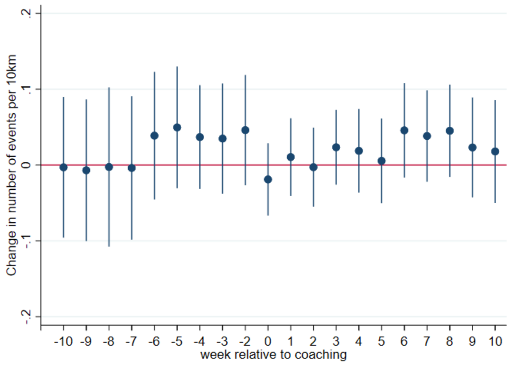
Figure 4: Dynamic Treatment Effects of In-Person Coaching [IW estimates]



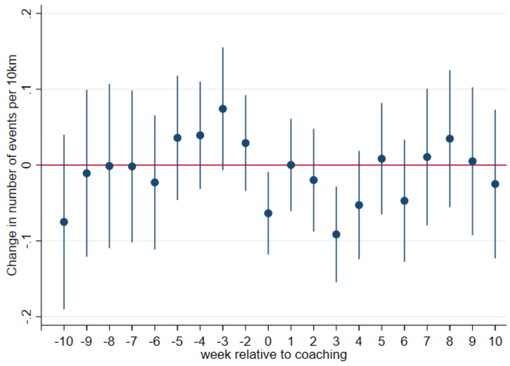
(a) Fuel economy



(b) Acceleration



(c) Braking



(d) Cornering

Notes: Sun and Abraham (2021) dynamic treatment estimates $\hat{\nu}_\ell$ that correct for possible heterogeneous treatment effects. The week of coaching itself is point 0 on the x-axis. The week before coaching is the omitted category. The vertical spikes indicate 95% confidence intervals. The dependent variable fuel economy is measured in liters/100km, and acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver. Coaches themselves are excluded from the sample. The considered time period is the period with complete logs available from all coaches (01/01/2015-30/04/2016).

Online Appendix [Not for Publication]

A Proofs

Derivation of Result 1

Driver d's decision problem is to maximize expected surplus:

$$S_d(e; k, t, t', c) = B(e) - \frac{C(e)}{q(t, t'; k, \gamma)} - N(e; \phi_{d,c}(t; t')), \text{ with } k > 1, \tag{A.1}$$

with e the chosen effort level, $B(e)$ the benefits of effort (with $B' > 0$ and $B'' < 0$), $C(e)/q(t, t'; k)$ the cost of effort (with $C', C'' > 0$) adjusted for worker quality $q(t, t'; k, \gamma) \equiv k + m(t; t')k^\gamma$, and $N(\cdot)$ (with $N' < 0$ and $N'' \geq 0$) the experienced effort norm. In the function for worker quality, k ($k > 1$) denotes the default worker quality and the function $m(t, t') \geq 0$ reflects how coaching at time t' affects worker quality at time t. This function is specified in equation (2).

The first-order condition of $S(e)$ with respect to e is:

$$\frac{\partial S(e)}{\partial e} = B'(e) - \frac{C'(e)}{k + m(t; t')k^\gamma} - N'(e) = 0. \tag{A.2}$$

To have an interior solution, the second-order condition

$$\frac{\partial^2 S(e)}{\partial e^2} = B''(e) - \frac{C''(e)}{k + m(t; t')k^\gamma} - N''(e) \tag{A.3}$$

must be negative, which holds because of the functional form assumptions imposed on $B()$, $C()$ and $N()$.

For a given constant $k > 1$, the optimal effort level is an implicit function of m, $e = g(m)$, where we omit the dependence of $m(\cdot)$ on t and t' for ease of exposition. Hence, the first-order condition can be written as:

$$\frac{\partial S(e)}{\partial m} = 0 \Rightarrow \frac{\partial S(e)}{\partial g}\frac{dg}{dm} + \frac{\partial S(e)}{\partial m} = 0 \Rightarrow \frac{dg}{dm} = \frac{de}{dm} = -\frac{\partial S(e)/\partial m}{\partial S(e)/\partial e}$$

or (see e.g Varian (1992), p. 492),

$$\frac{de}{dm} = -\frac{\partial^2 S(e)/\partial m \partial e}{\partial^2 S(e)/\partial e^2}.$$

Because $\frac{d}{dm}\left(\frac{C(e)}{k+m(t;t')k^\gamma}\right) = -\frac{C(e)k^\gamma}{(k+m(t;t')k^\gamma)^2}$, such that $\partial^2 S/\partial m \partial e = \frac{C'(e)k^\gamma}{(k+m(t;t')k^\gamma)^2} > 0$ for $k \geq 1$, and because the denominator is the second derivative of the objective function which must be negative for the solution to be an interior maximum, it follows that

$$\frac{de}{dm} > 0.$$

From this and from equation (2), it follows that when coaching has an immediate impact on worker quality ($b > 0$), $m(t; t')$ jumps upward at time $t = t'$ and hence increases the driver's optimal effort level; plus that when the effect is persistent ($\alpha > 0$), $m(t_2; t') > m(t_1; t')$ for $t_1 < t' < t_2$. This proves Result 1. $\square$

Derivation of Result 2

$$\frac{\partial^2 q}{\partial k \partial m} = \frac{\partial}{\partial k}\left(\frac{\partial q}{\partial m}\right) = \frac{\partial}{\partial k}\left(k^{\gamma}\right) = \gamma k^{\gamma-1} \Rightarrow \text{sign}\left(\frac{\partial^2 q}{\partial k \partial m}\right) = \text{sign}(\gamma).$$

In other words, for $\gamma > 0$ (< 0), the effect of coaching on worker quality is increasing (decreasing) in initial worker quality. □

Derivation of Result 3

The optimal effort level is an implicit function of ϕ, $e = h(\phi)$, where for ease of exposition, we omit the indices and dependence on t and t' in the parameter ϕ that denotes the salience of the effort norm. Hence, the first-order condition can be written as:

$$\frac{\partial S(e)}{\partial \phi} = 0 \Rightarrow \frac{\partial S(e)}{\partial h}\frac{dh}{d\phi} + \frac{\partial S(e)}{\partial \phi} = 0 \Rightarrow \frac{dh}{d\phi} = \frac{de}{d\phi} = -\frac{\partial S(e)/\partial \phi}{\partial S(e)/\partial e}$$

or

$$\frac{de}{d\phi} = -\frac{\partial^2 S(e)/\partial \phi \partial e}{\partial^2 S(e)/\partial e^2}.$$

We have assumed that the marginal effect of effort is increasing in the salience of the effort norm, $\partial^2 N(e;\phi)/\partial \phi \partial e < 0$, hence

$$\frac{\partial^2 S(e)}{\partial \phi \partial e} = -\frac{\partial^2 N(e;\phi)}{\partial \phi \partial e} > 0.$$

Because $-\partial^2 S(e)/\partial \phi \partial e < 0$ and because the denominator is the second derivative of the objective function which must be negative for the solution to be an interior maximum, it follows that

$$\frac{de}{d\phi} > 0.$$

□

B Worker productivity measures and scope for improvement

This section gives details on the definition and construction of the pre-experimental period worker quality rankings and the estimation of the residual standard deviation that is used to gauge the potential scope for improvement.

Pre-Experimental Period Worker Productivity Rankings We start with the trip-level sample that covers the period for which we have complete coach logs (April 30, 2016) and include per outcome all drivers i with $n_i \geq 25$ observations for that outcome. Let the outcome variable of interest, Y_{it} (fuel economy or ABC), indexed by driver (i) and trip (t), be given by:

$$Y_{it} = \mathbf{X_{it}} \cdot \beta + \nu_{it}, \tag{A.4}$$

with

$$\nu_{it} = \mu_i + \epsilon_{it}, \tag{A.5}$$

In this specification, the vector $\mathbf{X_{it}}$ includes all observable determinants of driver performance, day, and route fixed effects. The disturbance ν_{it} is decomposed into a driver fixed effect μ_i which absorbs all time-invariant unobservables at the driver level and idiosyncratic driver-level variation ϵ_{it}. Especially for the ABC-dimensions our baseline period before treatment is too small for a reliable estimation of the residual variation. Hence, we use for each driver all observations before the date of coaching to compute drivers' residual outcomes.[39]

We extract the driver fixed effects, which can be interpreted as the driving-context-adjusted estimates of worker quality for an outcome dimension. However, these estimates $\hat{\mu}_i$ may suffer from measurement error which cause attenuation bias in our analysis. We correct for this bias using an Empirical Bayes (EB) shrinkage procedure similar to Chandra et al. (2016, Online Appendix C).

The estimated driver quality $\hat{\mu}_i$ can be decomposed into the 'true' worker quality μ_i plus an disturbance η_i:

$$\hat{\mu}_i = \mu_i + \eta_i.$$

In line with the EB procedure outlined in Morris (1983), we assume that the estimated worker qualities are independent random draws from a normal distribution centered around the true quality, that is:

$$\hat{\mu}_i | \mu_i, \sigma^2_{\eta,i} \sim N(\mu_i, \sigma^2_{\eta,i}),$$

with $\sigma^2_{\eta,i}$ the variance of the measurement error of the estimate – which ceteris paribus will be decreasing in the number of observations n_i. And we make the assumption that true worker quality is independent normally distributed:

$$\mu_i | \mathbf{z_i}, \gamma, \sigma^2 \sim N(\mathbf{z_i'}\gamma, \sigma^2),$$

with σ^2 the underlying variance. This is the prior distribution.

[39] This approach is similar to Chetty, Friedman and Rockoff (2014) who use all observations to maximize statistical precision. We have sufficient pre-coaching observations and prefer, for reasons of inference, not to include observations once a driver has received coaching.

The posterior distribution of μ_i conditional of estimated quality is (cf. Morris (1983, (1.7)))

$$\mu_i | \hat{\mu}_i, \mathbf{z_i}, \gamma, \sigma^2, \sigma^2_{\eta,i} \sim N(\mu_i^{EB}, \sigma^2_{\eta,i}(1 - b_i)), \tag{A.6}$$

with μ_i^{EB} denoting the EB-adjusted quality, given by:

$$\mu_i^{EB} = (1 - b_i)\hat{\mu}_i + b_i \mathbf{z_i}'\gamma, \tag{A.7}$$

and

$$b_i = \frac{\sigma^2_{\eta,i}}{\sigma^2_{\eta,i} + \sigma^2}. \tag{A.8}$$

That is, the EB-adjusted estimate of worker quality is a weighted average of the estimated quality and the prior mean quality with the weight $b_i \in [0, 1]$ the attenuation factor or shrinkage coefficient. The higher the variance $\sigma^2_{\eta,i}$ of the estimated quality relative to the variance σ^2 of the true quality, the less weight will be given to the estimate.

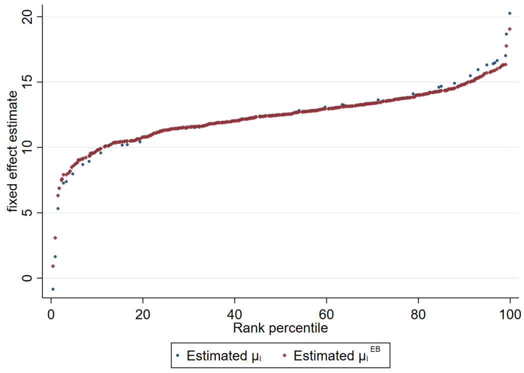
In the implementation, for all i, the $\sigma^2_{\eta,i}$'s are estimated by squaring the standard error of the fixed effect estimates $\hat{\mu}_i$ that result from the trip level regression (A.4)-(A.5), clustering the standard errors at the driver level. We will assume that the true underlying mean quality $\mathbf{z_i}'\gamma$ is equal to a base location fixed effect, i. e. $\mathbf{z_i}'\gamma = \tau_\ell$ with ℓ an index of all locations $\ell = 1, \ldots, L$ and τ_ℓ the fixed effect for location ℓ. We then perform the EB-procedure, giving estimates $\hat{\tau}_\ell$ of the underlying location mean qualities and within-location variance $\hat{\sigma}^2$.[40] The EB-adjusted quality measures μ_i^{EB} attenuate the estimated qualities to the estimated underlying mean quality at driver i's base location. Figure B.1 shows per driver and outcome dimension the uncorrected and EB-adjusted fixed effect estimates.

To estimate the standard deviation of the productivity metrics in Table 2, we do not simply use the standard deviation of the μ_i^{EB}'s – Chandra et al. (2016) warn that this would under-estimate dispersion – but use the formula they give for $\hat{\sigma}\xi^2$ ($\hat{\zeta}^2$ in their notation) in Online Appendix C.1, equation (A4) and take the root.

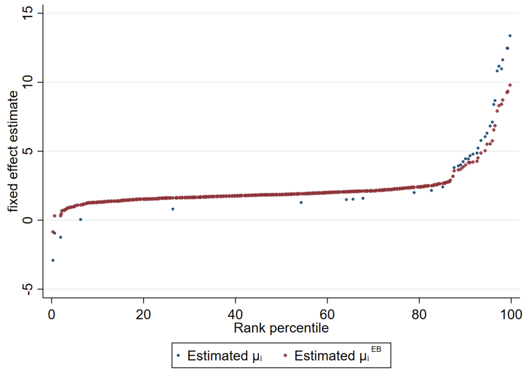
Other than Chandra et al. (2016), we report in Table 2 the raw correlations unadjusted for measurement error. Part of the correlation may stem from the fact that the same sample of drivers is used for estimating the various productivity metrics. In Chandra et al. (2016) this is a source of bias but our objective is exactly to show that drivers who perform well one on metric are more likely to perform well on the other metrics. A second source of bias is when one of the quality metrics is measured with error. This will bias their correlation downward. Hence the correlations we present should be considered a lower bound on the true correlation.

[40]We implement the EB-procedure in Stata using the `ebayes.do` program developed by Adam Sacarny.

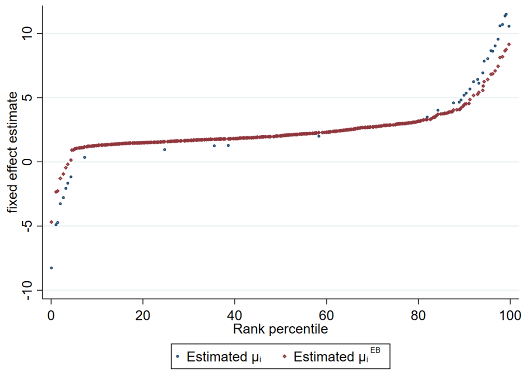
Figure B.1: Driver Fixed Effect Estimates



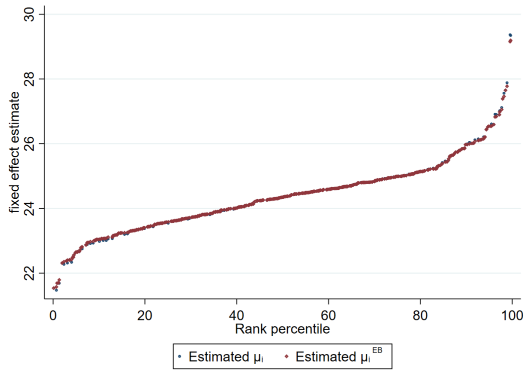
(a) Acceleration



(b) Braking



(c) Cornering



(d) Fuel economy

Notes: The figures show for fuel economy and each ABC comfort dimension the driver fixed effect estimates (A.5) and the Empirical-Bayes-adjusted driver fixed effect estimates (A.7), with drivers ranked from to the lowest to highest EB-adjusted fixed effect estimate. The dependent variable fuel economy is measured in liters/100km, and acceleration, braking and cornering as the number of events per 10 kilometers. Shown are the driver estimated average productivity measures with the driving condition adjustments set at their mean values.

C Determinants of Fuel Economy and ABC outcomes

Table C.1: Determinants of Fuel Economy

Dependent variable:		Fuel economy (in liters per 100km)				
		(1)	(2)	(3)	(4)	(5)
Rush hour 7-10am				-0.262***	-0.276***	-0.160***
				(0.035)	(0.035)	(0.019)
Rush hour 4-7pm				0.274***	0.296***	0.206***
				(0.037)	(0.036)	(0.020)
Non-scheduled trip				0.097	0.108	0.053
				(0.104)	(0.101)	(0.046)
No. of stops per km.				0.816***	0.855***	0.840***
				(0.102)	(0.104)	(0.096)
Trip length (in km.)				-0.021***	-0.023***	-0.023***
				(0.001)	(0.001)	(0.001)
Ln(No. of passengers)				1.299***	1.322***	1.318***
				(0.018)	(0.017)	(0.015)
Punctuality				0.020***	0.015***	0.048***
				(0.005)	(0.005)	(0.003)
Bus type:	VDL 10m	-0.924***	-0.926***	-0.878***	-0.878***	-0.872***
		(0.036)	(0.036)	(0.033)	(0.032)	(0.022)
	VDL 14m	9.208***	9.319***	9.319***	9.521***	9.307***
		(0.285)	(0.272)	(0.230)	(0.222)	(0.213)
	Intouro	4.869***	4.908***	4.490***	4.501***	4.474***
		(0.362)	(0.365)	(0.366)	(0.345)	(0.222)
Constant		23.926***	23.656***	20.412***	21.950***	22.779***
		(0.062)	(0.066)	(0.159)	(0.316)	(0.295)
R^2		.304	.317	.43	.448	.53
Number of trip-level observations		279501	279501	279501	279501	279501
Weather dummies		No	Yes	Yes	No	No
Driver fixed effects		No	No	No	No	Yes
Day fixed effects		No	No	No	Yes	Yes
Route fixed effects		No	No	Yes	Yes	Yes

Notes: Dependent variable: Fuel economy in liters/100km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Trip punctuality is the difference in minutes between actual and scheduled driving time. Weather data are collected from a weather station located in the regional capital and are maintained by the Royal Netherlands Meteorological Institute (KNMI). Standard errors are clustered by driver.
$^{***}(^{**},^{*})$: statistically different from zero at the 1%-level (5%-level, 10%-level).

Table C.2: Determinants of Acceleration

Dependent variable:		Acceleration				
		(1)	(2)	(3)	(4)	(5)
Rush hour 7-10am				-0.375**	-0.430***	-0.224**
				(0.156)	(0.157)	(0.111)
Rush hour 4-7pm				0.712***	0.772***	0.623***
				(0.137)	(0.134)	(0.086)
Non-scheduled trip				0.167	0.220	0.370***
				(0.313)	(0.319)	(0.130)
No. of stops per km.				3.147***	3.106***	3.275***
				(0.373)	(0.353)	(0.301)
Trip length (in km.)				-0.021***	-0.020***	-0.018***
				(0.003)	(0.002)	(0.002)
Ln(No. of passengers)				1.450***	1.376***	1.318***
				(0.064)	(0.057)	(0.053)
Punctuality				-0.055***	-0.046***	0.027***
				(0.012)	(0.011)	(0.008)
Bus type:	VDL 10m	0.076	0.095	0.149	0.215**	0.177**
		(0.109)	(0.103)	(0.102)	(0.103)	(0.070)
	VDL 14m	-0.569**	-0.643**	-0.739**	-0.487	-0.334
		(0.289)	(0.283)	(0.310)	(0.330)	(0.314)
	Intouro	-0.925**	-0.095	-0.507	0.018	0.548
		(0.411)	(0.439)	(0.409)	(0.371)	(0.507)
Constant		10.679***	11.567***	3.982***	1.296***	2.752***
		(0.155)	(0.146)	(0.473)	(0.365)	(0.308)
R^2		.492	.502	.534	.549	.633
Number of trip-level observations		118051	118051	118051	118051	118051
Weather dummies		No	Yes	Yes	No	No
Driver fixed effects		No	No	No	No	Yes
Day fixed effects		No	No	No	Yes	Yes
Route fixed effects		No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of acceleration events per 10 km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Standard errors are clustered by driver.
***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

Table C.3: Determinants of Braking

Dependent variable:		**Braking**				
		(1)	(2)	(3)	(4)	(5)
Rush hour 7-10am				0.038	-0.001	0.053
				(0.061)	(0.055)	(0.039)
Rush hour 4-7pm				0.148**	0.195***	0.169***
				(0.065)	(0.062)	(0.030)
Non-scheduled trip				-0.145	-0.157	-0.120
				(0.112)	(0.108)	(0.087)
No. of stops per km.				0.475***	0.469***	0.660***
				(0.180)	(0.160)	(0.084)
Trip length (in km.)				-0.005***	-0.004***	-0.003***
				(0.001)	(0.001)	(0.000)
Ln(No. of passengers)				0.240***	0.187***	0.190***
				(0.032)	(0.025)	(0.023)
Punctuality				-0.019***	-0.011**	0.008***
				(0.005)	(0.005)	(0.003)
Bus type:	VDL 10m	-0.004	0.009	0.012	0.052**	0.030
		(0.027)	(0.024)	(0.024)	(0.025)	(0.019)
	VDL 14m	0.244***	0.165	0.145	0.297**	0.141
		(0.077)	(0.111)	(0.114)	(0.121)	(0.131)
	Intouro	3.063***	3.573***	3.462***	3.842***	3.882***
		(0.550)	(0.549)	(0.541)	(0.540)	(0.520)
Constant		1.288***	1.848***	0.651**	-0.676***	-0.486***
		(0.096)	(0.097)	(0.302)	(0.173)	(0.144)
R^2		.377	.401	.406	.448	.601
Number of trip-level observations		118051	118051	118051	118051	118051
Weather dummies		No	Yes	Yes	No	No
Driver fixed effects		No	No	No	No	Yes
Day fixed effects		No	No	No	Yes	Yes
Route fixed effects		No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of braking events per 10 km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Standard errors are clustered by driver.
***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

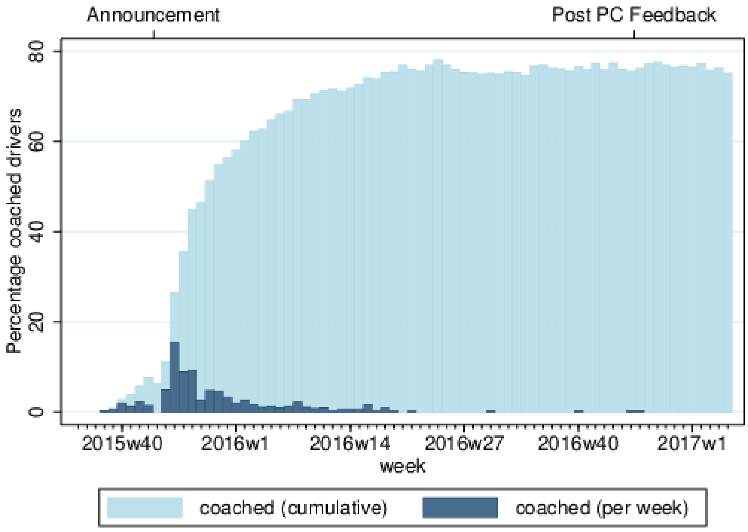
Table C.4: Determinants of Cornering

Dependent variable:		**Cornering**				
		(1)	(2)	(3)	(4)	(5)
Rush hour 7-10am				-0.155**	-0.161**	-0.051
				(0.077)	(0.077)	(0.031)
Rush hour 4-7pm				0.097	0.112	0.013
				(0.089)	(0.089)	(0.022)
Non-scheduled trip				-0.012	-0.000	-0.147
				(0.147)	(0.143)	(0.106)
No. of stops per km.				-0.439***	-0.443***	-0.401***
				(0.168)	(0.158)	(0.104)
Trip length (in km.)				-0.005***	-0.004***	-0.005***
				(0.001)	(0.001)	(0.001)
Ln(No. of passengers)				0.097***	0.055***	0.042***
				(0.021)	(0.018)	(0.015)
Punctuality				-0.069***	-0.065***	-0.025***
				(0.007)	(0.007)	(0.003)
Bus type:	VDL 10m	0.016	0.021	0.001	0.034	0.013
		(0.034)	(0.033)	(0.032)	(0.033)	(0.018)
	VDL 14m	-0.184**	-0.166*	-0.169*	-0.191	-0.459***
		(0.074)	(0.096)	(0.102)	(0.139)	(0.145)
	Intouro	-0.386***	-0.165	-0.217	-0.130	0.003
		(0.123)	(0.134)	(0.147)	(0.154)	(0.094)
Constant		1.721***	2.030***	2.426***	0.892***	0.035
		(0.111)	(0.115)	(0.228)	(0.181)	(0.127)
R^2		.567	.571	.574	.584	.736
Number of trip-level observations		118051	118051	118051	118051	118051
Weather dummies		No	Yes	Yes	No	No
Driver fixed effects		No	No	No	No	Yes
Day fixed effects		No	No	No	Yes	Yes
Route fixed effects		No	No	Yes	Yes	Yes

Notes: Dependent variable: Number of cornering events per 10 km. Default categories: Scheduled trips in rural areas outside rush hours completed with a VDL 12m bus. Standard errors are clustered by driver.
***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

D Timing and Quasi-Random Assignment of Coaching sessions

Figure D.2: Time of First Coaching

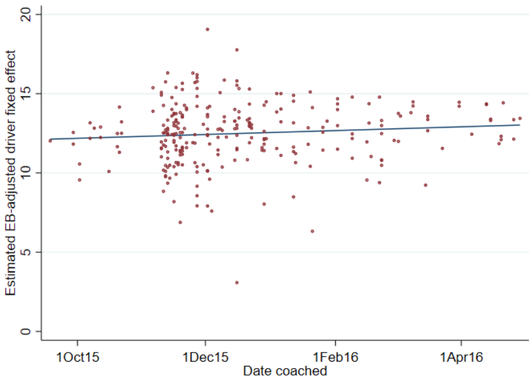


Notes: Moment of first coaching for drivers. Dark blue bars indicate the drivers who received their first coaching during a specific week as a share of the total number of drivers operating during that week. The light blue bars depict the cumulative share of coached drivers operating during a week. Feedback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

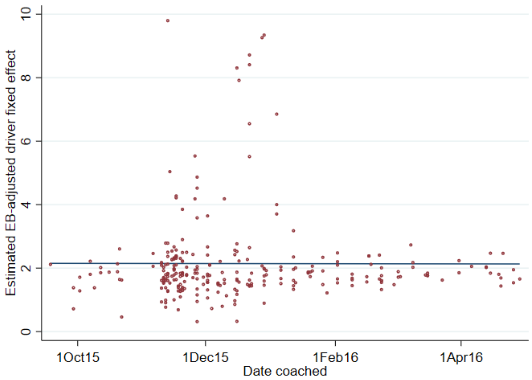
Two things in Figure D.2 related to the coaching program deserve some further explanation. First, the cumulative share of coached drivers operating during a week is more or less flat after April 2016. We have complete coach logs for the period till 30 April, 2016. Some coaches indicated that they no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

Figure D.3 plots for the four outcomes the relation between a driver's EB-adjusted pre-coaching productivity metric and the time the driver receives the first coaching session. Under quasi-random assignment of coaching sessions, a driver's past performance is unrelated to the date at which the driver is selected for coaching. The figure also plots the linear trend lines from a linear regression of $\hat{\mu}_i^{EB}$ on the date of coaching. For none of the outcomes we find a statistically significant increasing (higher productivity drivers tend to be coached first) or decreasing trend.

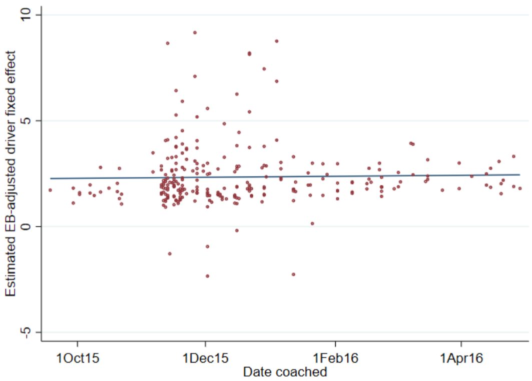
Figure D.3: Driver Specific Estimated EB-adjusted Fixed Effects and Time of Coaching



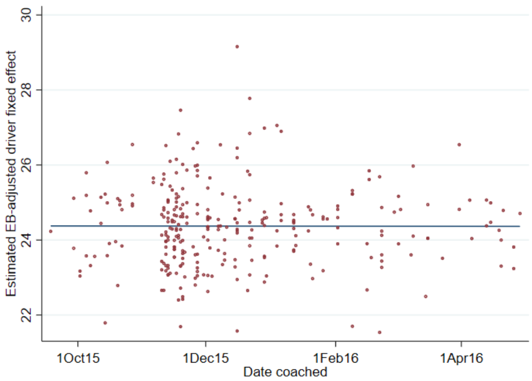
(a) Acceleration (p-value trend = 0.124)



(b) Braking (p-value trend = 0.964)



(c) Cornering (p-value trend = 0.694)



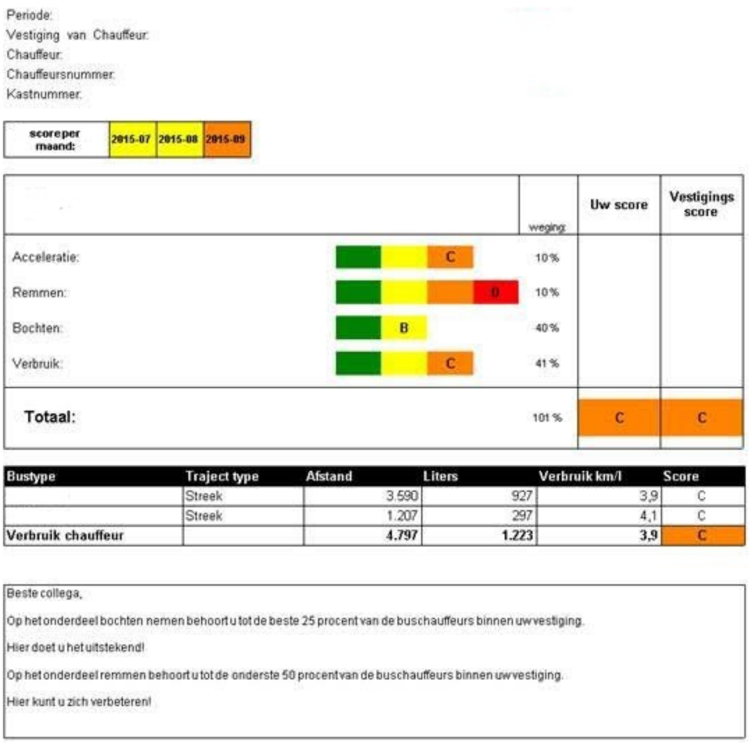
(d) Fuel economy (p-value trend = 0.970)

Notes: Plotted are the EB-adjusted fixed effects estimates $\hat{\mu}_i^{EB}$ of drivers who receive their first coaching session at a particular date and a linear trend line from a linear regression of $\hat{\mu}_i^{EB}$ on the date of coaching. Shown are the driver estimated average productivity measures with the driving condition adjustments set at their mean values.

E Sample Feedback Report

Figure E.4 reproduces a specimen of the feedback report drivers received once a month between December 2015 - November 2016.

Figure E.4: 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!

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

F Construction Final Sample

The initial sample consists of 1,278,913 trip-level observations. Table F.5 summarizes the steps taken to arrive at the final sample that is used in our analysis. The absence of a bus identifier has been the most important reason to discard observations (22.7% of the initial sample). Without such an identifier, it is not possible to link the trip to the outcome data of interest (fuel economy and ABC dimensions). The missing of identifiers is bus specific and is caused for technical reasons. Besides the standard set of identified buses the company also frequently uses buses that do not transfer data to the data warehouse. This can be older buses that are occasionally used to fill gaps, buses that are transferred from other concession areas or buses rented or bought from other bus companies. Observations are also dropped when there are technical mismatches between the bus type and EOBR data (about 5%) or when there are between-trip inconsistencies at the driver level (3%). We also exclude extreme outcomes regarding punctuality or outcomes that suggest a temporary technical recording problem in the EOBR (< 0.5 percent). These comprise observations of fuel economy being less than one or more than eight (1,259 obs; 0.10%), a difference of more than one hour between actual and planned driving time (156 obs; 0.01%) and outcomes that are more than five standard deviations above the means of the ABC dimensions (4,003 obs; 0.31%).

Table F.5: 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)	0.50
No bus identifier	(290,737)	22.73
Bus type not eligible for EOBR	(34,870)	2.73
Error message from EOBR	(31,118)	2.46
Within-driver obs. with the same departure date/time	(37,575)	2.94
Very short rides (less than 1 kilometer)	(29,588)	2.31
Punctuality: more than 1 hour	(156)	0.01
Unreasonable outcomes of dependent variables:		
- Fuel economy: less than 12.5 or more than 100	(1,259)	0.10
- ABC dimensions: more than 5 SDs above the mean	(4,003)	0.31
	842,845	

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

The company uses different bus types. For the trips in the final sample, Table F.6 shows which bus type is used. The table reveals that the vast majority of trips (> 63%) is completed with a VDL bus. More specifically, over half of the trips are completed with the VDL 12 meter bus. The IRIS bus is used in about 30% of all trips. Most of these IRIS buses run on natural gas, which implies that no fuel economy score is recorded for these trips. These gas buses are almost exclusively used in the province's capital because of the lower CO_2 emissions of natural gas compared to diesel. A small minority (<10%) of trips is completed with a third bus type, the Mercedes Intouro.

Table F.6: Fleet Information

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

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

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

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

Notes: The number of trips for which the ABC events are recorded is lower than the total number of trips mentioned in the table, because recording of ABC events commenced only in Sept. 2015. This explains the difference between the number of observations mentioned in this table and those in the regression tables of Section 5.

G Intermediate Changes in ABC Settings

During the period of data collection, the company changed some of the threshold settings for the ABC dimensions. An increase (decrease) in the threshold has the effect of reducing (increasing) the number of recorded events. The company has provided a detailed list of when which threshold has been changed on which bus type(s). This list is presented in Table G.8.

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

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

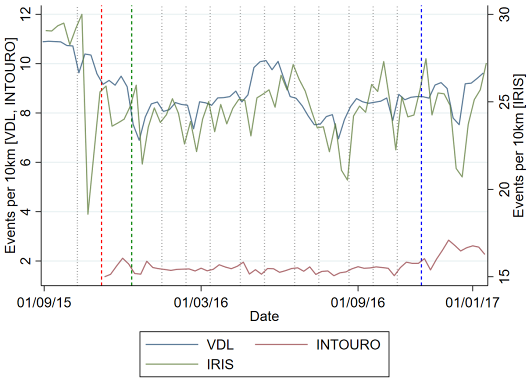
The Intouro buses experienced a number of recalibrations, but all before the date the first feedback report was received by the drivers and mostly comprised only two to four buses out of a total of 29 Intouro buses. Because calibration for these buses was late, we have no records for acceleration and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

Other than the ABC dimensions, the fuel economy records of both the VDL and Intouro buses cover the entire period from January 2015 till January 2017 and have not been subject to any change in measurement. For completeness, Figures G.5d shows the weekly average fuel economy for both the VDL and Intouro buses.

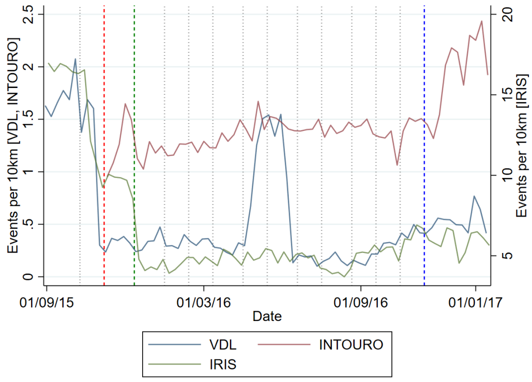
Table G.8: Change in Threshold Settings

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

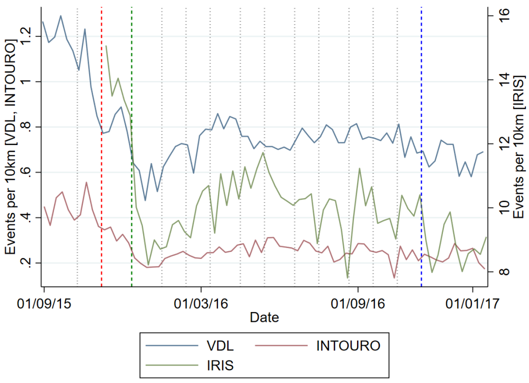
Figure G.5: Development Over Time in ABC Dimensions (No. Events) and Fuel Economy (liters/100km) – Weekly Averages by Bus Type



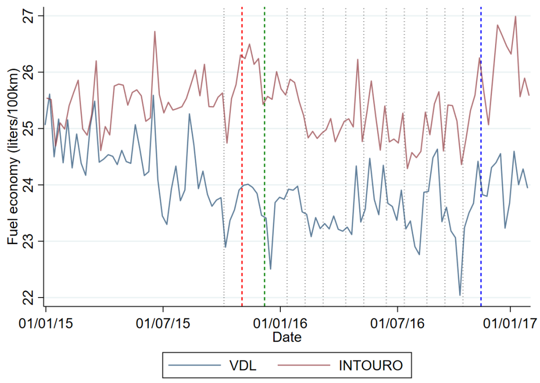
(a) Acceleration



(b) Braking



(c) Cornering

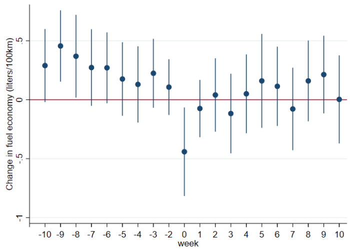


(d) Fuel economy

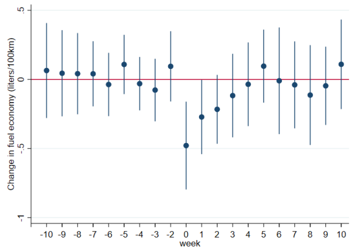
Notes: Averages are calculated based on all trips in the analysis set. The vertical dashed lines indicate: the launch date of the EcoManager program [red, 09/11/2015], the distribution of the first feedback report [green, 15/12/2015] and the distribution of the report with the final notification message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

H Temporal Effects In-Person Coaching at Coach Level

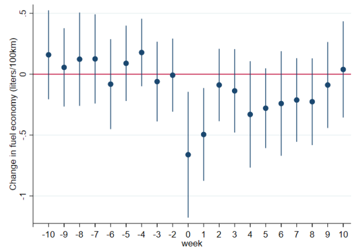
Figure H.6: Dynamic Treatment Effects In-Person Coaching at Coach Level: Fuel Economy



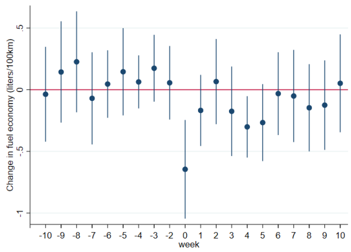
(a) Coach # 1



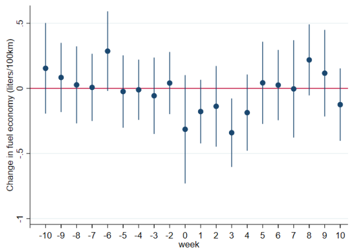
(b) Coach # 2



(c) Coach # 3



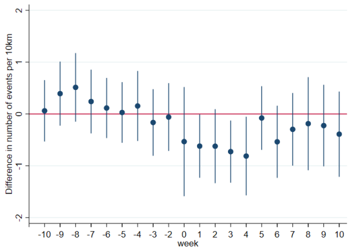
(d) Coach # 4



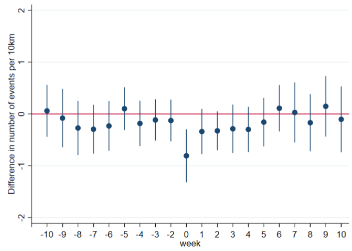
(e) Coach # 5

Note: For one coach there are no estimates because of insufficient observations because this coach operates in an (urban) area with IRIS buses that do not record fuel economy.

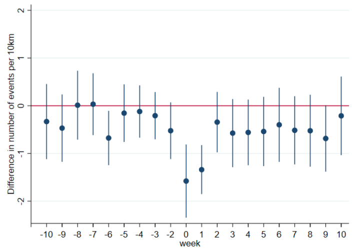
Figure H.7: Dynamic Treatment Effects In-Person Coaching at Coach Level: Acceleration



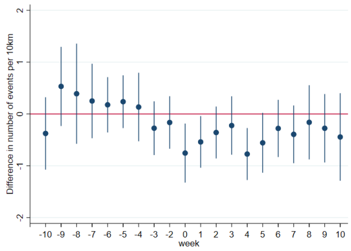
(a) Coach # 1



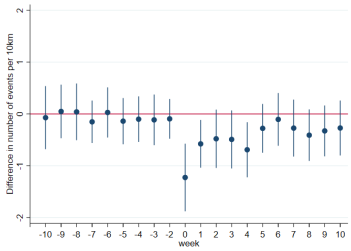
(b) Coach # 2



(c) Coach # 3

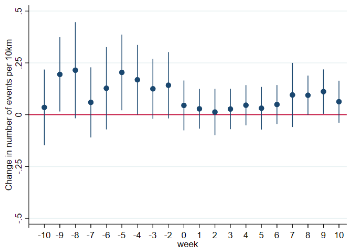


(d) Coach # 4

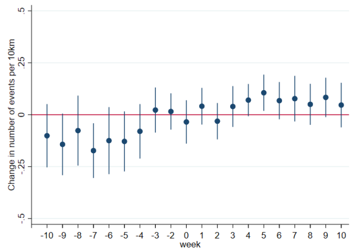


(e) Coach # 5

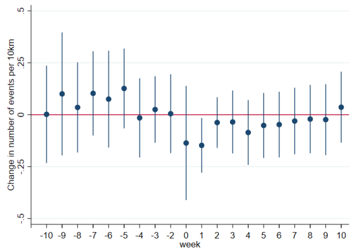
Figure H.8: Dynamic Treatment Effects In-Person Coaching at Coach Level: Braking



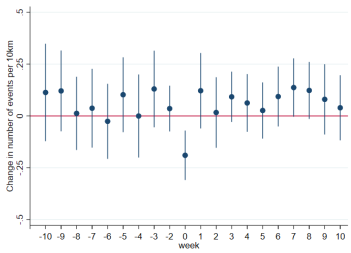
(a) Coach # 1



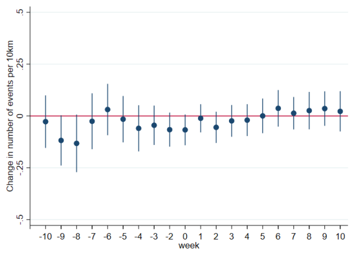
(b) Coach # 2



(c) Coach # 3

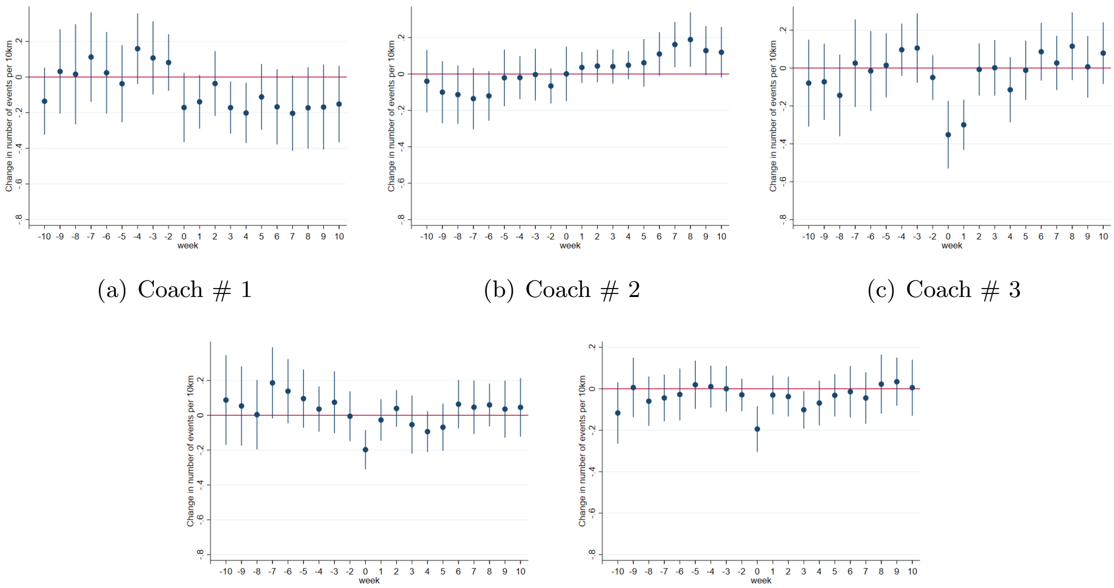


(d) Coach # 4



(e) Coach # 5

Figure H.9: Dynamic Treatment Effects In-Person Coaching at Coach Level: Cornering



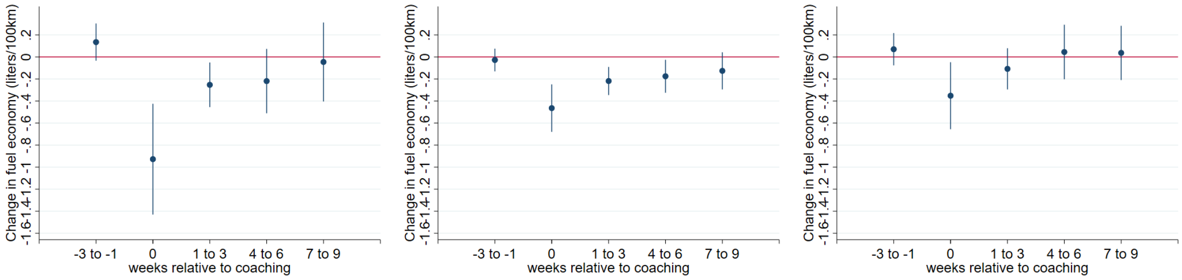
(a) Coach # 1 (b) Coach # 2 (c) Coach # 3

(d) Coach # 4 (e) Coach # 5

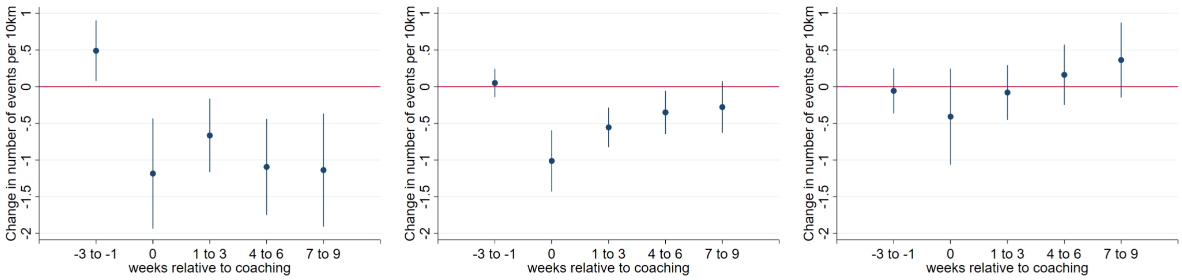
I Treatment Effect Heterogeneity Coaching

Figure I.10: Heterogeneous Treatment Effects of Coaching - Treatment Region

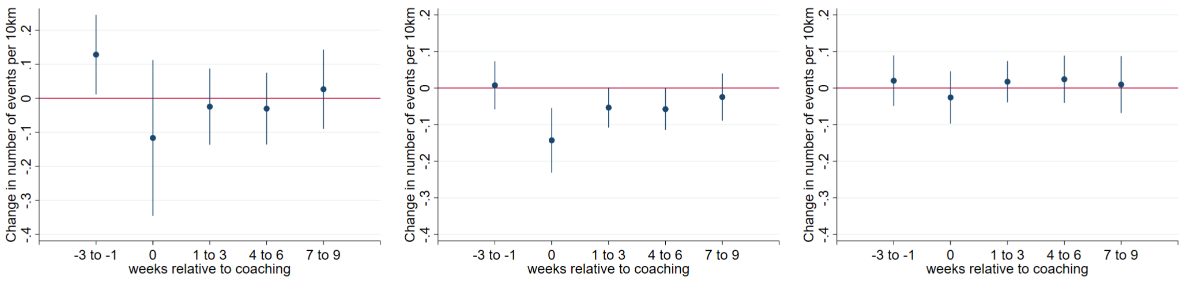
LOW Productivity $p(\hat{\mu}^{EB}) \geq 75$	MEDIUM Productivity $25 < p(\hat{\mu}^{EB}) < 75$	HIGH Productivity $p(\hat{\mu}^{EB}) \leq 25$



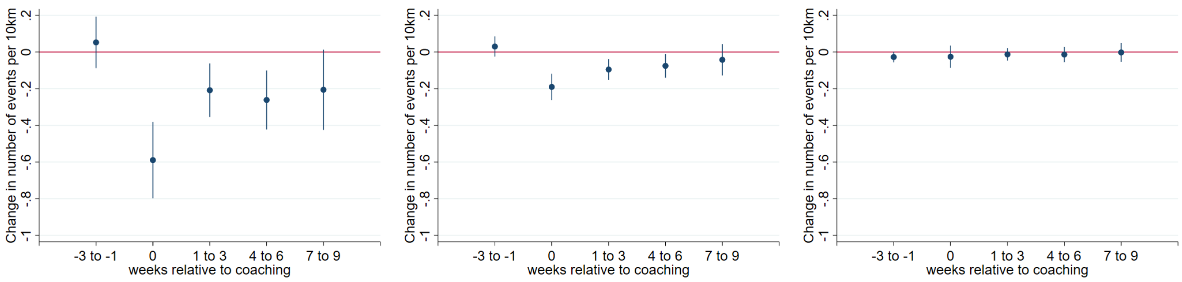
(a) Fuel economy



(b) Acceleration



(c) Braking

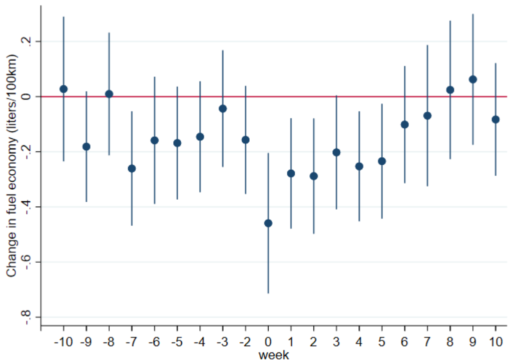


(d) Cornering

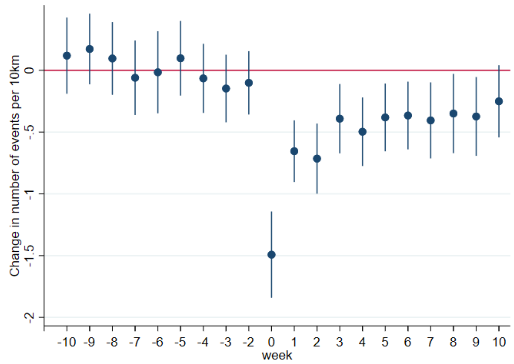
Notes: TWFE dynamic treatment estimates of equation (6) (with weeks replaced by three week periods) on subsamples that condition on the EB-adjusted pre-coaching worker productivity estimates $\hat{\mu}^{EB}$ for the outcome considered. The left panels show the treatment effects for drivers with LOW pre-coaching productivity (percentile rank $p(\hat{\mu}^{EB}) \geq 75$); the middle panels idem ($25 < p(\hat{\mu}^{EB}) < 75$); the right panels idem ($p(\hat{\mu}^{EB}) \leq 25$). The seven days before coaching are the omitted category. "Week 0" is the day coaching is received. The vertical spikes indicate 95% confidence intervals. The dependent variable fuel economy is measured in liters/100km, and acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver. Controls include: travel distance, number of passengers and bus stops, dummies for non-scheduled rides, bus types, morning and evening rush hours, and the interaction of bus type and day fixed effects. Coaches themselves are excluded. The considered time period: 01/01/2015-30/04/2016.

J Replication of In-Person Coaching Results in the Control Region

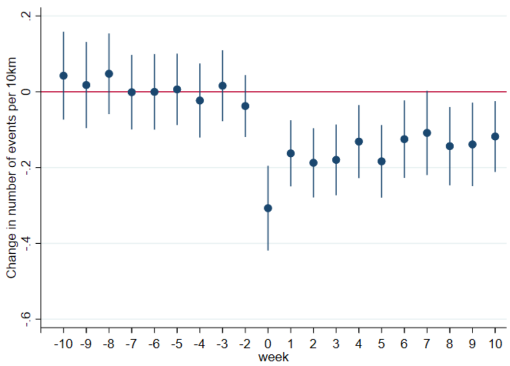
Figure J.11: Dynamic Treatment Effects of In-Person Coaching [TWFE] - Control Region



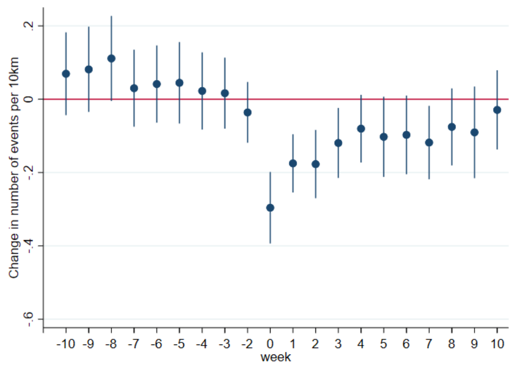
(a) Fuel economy



(b) Acceleration



(c) Braking



(d) Cornering

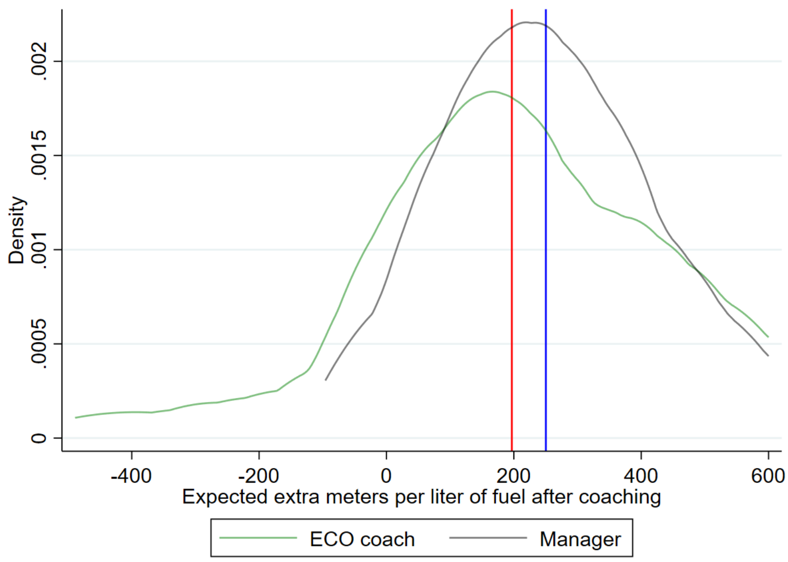
Notes: Conventional TWFE dynamic treatment estimates $\hat{\delta}_t$. Driving performance in the 10 weeks before and after coaching. The day of coaching itself is point 0 on the x-axis. The vertical spikes indicate 95% confidence intervals. The dependent variable fuel economy is measured in liters/100km,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

K Survey Among Employees About Management Views on In-Person Coaching

Table K.9: Responses to the Survey Question:
Should Coaching be Offered if the Costs Exceed the Fuel Savings Realized by a Coaching Session?

Answer	Percentage
Yes, but only if next to small fuel savings it also leads to improvements on other dimensions (such as the environment, passenger comfort, and personnel development)	82.7%
Yes, coaching should be offered, even if it only leads to small fuel savings	9.9%
No, coaching should only be offered if it leads to lower net costs	1.2%
Other	6.2%

Figure K.12: Expectations of Coaches and Managers:
How Many Extra Meters per Liter of Fuel Right After a Coaching Session?

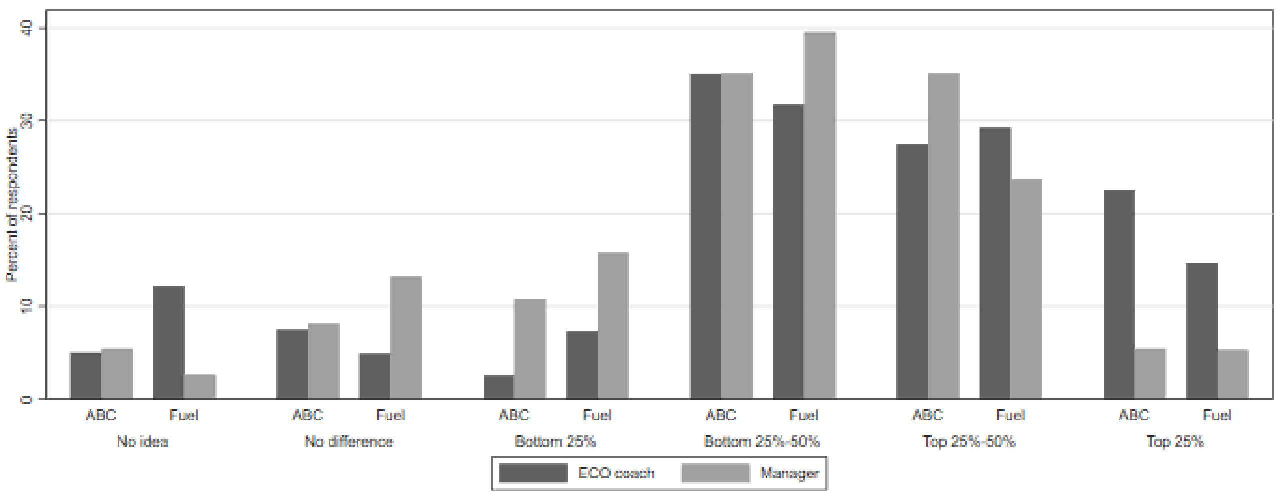


Note: The vertical red (blue) line is the average estimate of coaches (managers).

Table K.10: Responses to the Survey Question: Will Coaching Have a Lasting Effect on Fuel Consumption?

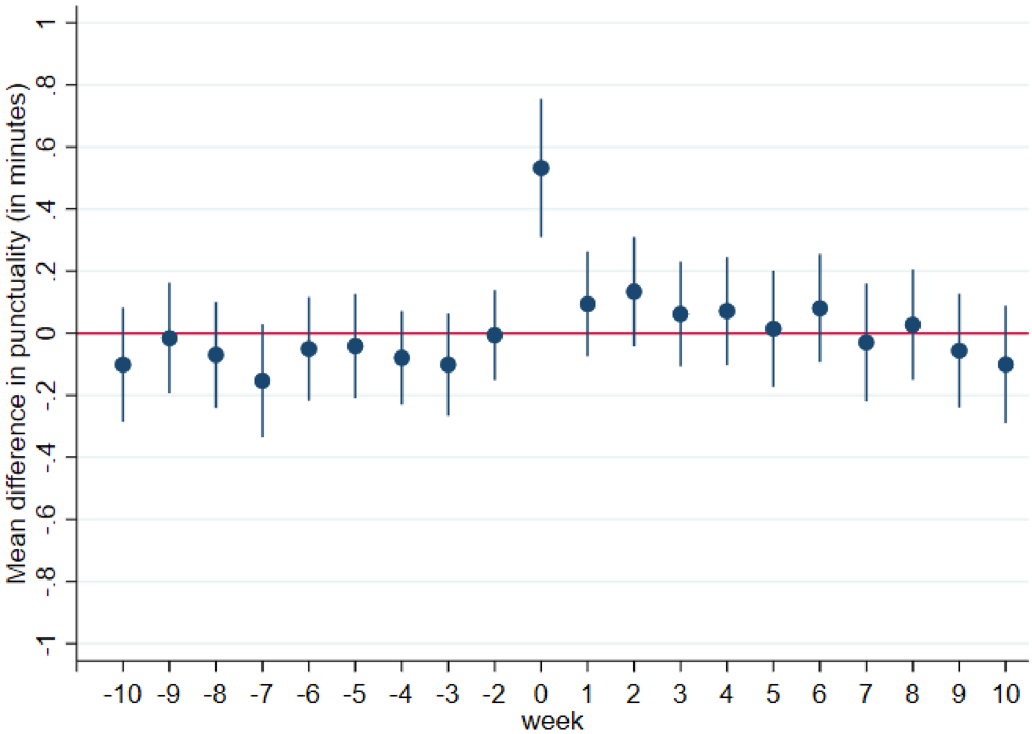
Answer	Coaches	Managers	Regional Directors
Yes	18 (41%)	18 (46%)	3 (60%)
No	24 (59%)	21 (54%)	2 (40%)

Figure K.13: Expectations of Coaches and Managers:
Which Group of Drivers Will Benefit the Most from Coaching?



L Further Robustness Checks

Figure L.14: Dynamic Treatment Effects of In-Person Coaching on Punctuality [TWFE]



Notes: Conventional TWFE dynamic treatment estimates $\hat{\delta}_t$. Punctuality in the 10 weeks before and after coaching. The day of coaching itself is point 0 on the x-axis. The vertical spikes indicate 95% confidence intervals. Punctuality is defined as the difference between the actual and scheduled trip length (in minutes) Standard errors are clustered by driver. Controls include: travel distance, number of passengers and bus stops, dummies for non-scheduled rides, bus types, morning and evening rush hours, and the interaction of bus type and day fixed effects. Coaches themselves are excluded from the sample.

Table L.11: Dynamic Treatment Effects In-Person Coaching on Driving Performance – Rush Hours Excluded as Control Variables

Dep. var.	**Fuel Economy**	**Acceleration**	**Braking**	**Cornering**
	(1)	(2)	(3)	(4)
	$\hat{\delta}_\tau$	$\hat{\delta}_\tau$	$\hat{\delta}_\tau$	$\hat{\delta}_\tau$
$\tau = -10$	0.077	-0.093	-0.009	-0.072
	(0.084)	(0.160)	(0.043)	(0.048)
-9	0.111	0.084	0.046	0.000
	(0.074)	(0.153)	(0.046)	(0.048)
-8	0.093	0.085	0.038	-0.037
	(0.074)	(0.153)	(0.047)	(0.049)
-7	0.037	-0.055	0.011	0.016
	(0.070)	(0.130)	(0.039)	(0.048)
-6	0.074	-0.030	0.032	-0.014
	(0.068)	(0.128)	(0.041)	(0.042)
-5	0.034	-0.061	0.072*	0.005
	(0.065)	(0.118)	(0.038)	(0.038)
-4	0.022	0.032	0.018	0.066*
	(0.060)	(0.126)	(0.037)	(0.034)
-3	0.038	-0.158	0.058*	0.040
	(0.068)	(0.123)	(0.031)	(0.037)
-2	0.041	-0.184	0.035	-0.007
	(0.061)	(0.118)	(0.030)	(0.027)
-1	0	0	0	0
	()	()	()	()
0	-0.525***	-1.074***	-0.081**	-0.190***
	(0.093)	(0.167)	(0.033)	(0.035)
1	-0.237***	-0.706***	-0.003	-0.103***
	(0.064)	(0.121)	(0.025)	(0.028)
2	-0.096	-0.503***	-0.032	-0.024
	(0.070)	(0.135)	(0.026)	(0.029)
3	-0.197***	-0.538***	0.008	-0.077***
	(0.072)	(0.136)	(0.024)	(0.029)
4	-0.172**	-0.699***	-0.001	-0.096***
	(0.077)	(0.144)	(0.025)	(0.031)
5	-0.079	-0.405***	0.010	-0.051
	(0.077)	(0.135)	(0.026)	(0.032)
6	-0.086	-0.384**	0.026	-0.021
	(0.083)	(0.151)	(0.027)	(0.037)
7	-0.116	-0.405**	0.038	-0.037
	(0.084)	(0.157)	(0.032)	(0.037)
8	-0.058	-0.434**	0.029	0.011
	(0.081)	(0.171)	(0.028)	(0.043)
9	0.001	-0.400**	0.046	-0.019
	(0.082)	(0.166)	(0.030)	(0.043)
10	0.005	-0.373**	0.035	-0.005
	(0.087)	(0.187)	(0.031)	(0.043)
Wald test				
leads $_{\tau=-10,\dots,-2}$	0.939	0.506	0.417	0.044
lags $_{\tau=0,\dots,10}$	<0.0001	<0.0001	0.061	<0.0001
Obs. (trips)	352,253	187,127	187,127	187,127
Nr. drivers	399	376	376	376
Controls	Yes	Yes	Yes	Yes

This table reports the dynamic treatment effects of in-person coaching on driving performance. TWFE estimates. $\hat{\delta}_0$ is the estimated effect at the day of coaching, $\hat{\delta}_1$ ($\hat{\delta}_{-1}$) the estimated effect 1-7 days after (before) coaching etc. The considered time period is 01/07/15-31/01/17. The dependent variable fuel economy is measured in liters/100km, acceleration, braking and cornering as the number of events per 10 kilometers. Standard errors are clustered by driver and shown in parentheses. The omitted category is the seven days (week) before coaching. Controls include: travel distance, route dummies, number of passengers and bus stops, and dummies for bus type, and non-scheduled rides.
***(**,*) : statistically different from zero at the 1%-level (5%-level, 10%-level).

List of research reports

2019001-EEF: Lugalla, I.M., J. Jacobs, and W. Westerman, Drivers of Women Entrepreneurs in Tourism in Tanzania: Capital, Goal Setting and Business Growth

2019002-EEF: Brock, E.O. de, On Incremental and Agile Development of (Information) Systems

2019003-OPERA: Laan, N. van der, R.H. Teunter, W. Romeijnders, and O.A. Kilic, The Data-driven Newsvendor Problem: Achieving On-target Service Levels.

2019004-EEF: Dijk, H., and J. Mierau, Mental Health over the Life Course: Evidence for a U-Shape?

2019005-EEF: Freriks, R.D., and J.O. Mierau, Heterogeneous Effects of School Resources on Child Mental Health Development: Evidence from the Netherlands.

2019006-OPERA: Broek, M.A.J. uit het, R.H. Teunter, B. de Jonge, J. Veldman, Joint Condition-based Maintenance and Condition-based Production Optimization.

2019007-OPERA: Broek, M.A.J. uit het, R.H. Teunter, B. de Jonge, J. Veldman, Joint Condition-based Maintenance and Load-sharing Optimization for Multi-unit Systems with Economic Dependency

2019008-EEF: Keller, J.T. G.H. Kuper, and M. Mulder, Competition under Regulation: Do Regulated Gas Transmission System Operators in Merged Markets Compete on Network Tariffs?

2019009-EEF: Hulshof, D. and M. Mulder, Renewable Energy Use as Environmental CSR Behavior and the Impact on Firm Profit

2019010-EEF: Boot, T., Confidence Regions for Averaging Estimators
2020001-OPERA: Foreest, N.D. van, and J. Wijngaard. On Proportionally Fair Solutions for the Divorced-Parents Problem

2020002-EEF: Niccodemi, G., R. Alessie, V. Angelini, J. Mierau, and T. Wansbeek. Refining Clustered Standard Errors with Few Clusters

2020003-I&O: Bogt, H. ter, Performance and other Accounting Information in the Public Sector: A Prominent Role in the Politicians' Control Tasks?

2020004-I&O: Fisch, C., M. Wyrwich, T.L. Nguyen, and J.H. Block, Historical Institutional Differences and Entrepreneurship: The Case of Socialist Legacy in Vietnam

2020005-I&O: Fritsch, M. and M. Wyrwich. Is Innovation (Increasingly) Concentrated in Large Cities? An Internatinal Comparison

2020006-GEM: Oosterhaven, J., Decomposing Economic Growth Decompositions.

2020007-I&O: Fritsch, M., M. Obschonka, F. Wahl, and M. Wyrwich. The Deep Imprint of Roman Sandals: Evidence of Long-lasting Effects of Roman Rule on Personality, Economic Performance, and Well-Being in Germany

2020008-EEF: Heijnen, P., On the Computation of Equilibrium in Discontinuous Economic Games

2020009-EEF: Romensen, G.J. and A.R. Soetevent, Improving Worker Productivity Through Tailored Performance Feedback: Field Experimental Evidence from Bus Drivers

2020010-EEF: Rao, Z., M. Groneck, and R. Alessie, Should I Stay or Should I Go? Intergenerational Transfers and Residential Choice. Evidence from China

2020011-EEF: Kwaak, C. van der, Unintended Consequences of Central Bank Lending in Financial Crises

2020012-EEF: Soetevent, A.R., Determinants choice set variation in demand estimation – with an application to the electric vehicle public charging market

2020013-EEF: Kwaak, C. van der, Old-Keynesianism in the New Keynesian model

2020014-EEF: Plaat, M. van der, Loan Sales and the Tyranny of Distance in U.S. Residential Mortgage Lending

2020015-I&O: Fritsch, M., and M. Wyrwich, Initial Conditions and Regional Performance in the Aftermath of Disruptive Shocks: The Case of East Germany after Socialism

2020016-OPERA: Laan, N. van der, and W. Romeijnders, A Converging Benders' Decomposition Algorithm for Two-stage Mixed-integer Recourse Models

2021001-OPERA: Baardman, L., K.J. Roodbergen, H.J. Carlo, and A.H. Schrotenboer, A Special Case of the Multiple Traveling Salesmen Problem in End-of-aisle Picking Systems

2021002-EEF: Wiese, R., and S. Eriksen, Willingness to Pay for Improved Public Education and Public Health Systems: The Role of Income Mobility Prospects.

2021003-EEF: Keller, J.T., G.H. Kuper, and M. Mulder, Challenging Natural Monopolies: Assessing Market Power of Gas Transmission System Operators for Cross-Border Capacity

2021004-EEF: Li, X., and M. Mulder, Value of Power-to-Gas as a Flexibililty Option in Integrated Electricity and Hydrogen Markets
2021005-GEM: Rozite, K., J.P.A.M. Jacobs, and D.J. Bezemer, Investor Sentiment and Business Investment

2021006-EEF: Spierdijk, L., and T. Wansbeek, Differencing as a Consistency Test for the Within Estimator

2021007-EEF: Katz, M., and C. van der Kwaak, To Bail-in or to Bailout: that's the (Macro) Question

2021008-EEF: Haan, M.A., N.E. Stoffers, and G.T.J. Zwart, Choosing Your Battles: Endogenous Multihoming and Platform Competition

2021009-I&O: Greve, M., M. Fritsch, and M. Wyrwich, Long-Term Decline of Regions and the Rise of Populism: The Case of Germany

2021010-MARK: Hirche, C.F., T.H.A. Bijmolt, and M.J. Gijsenberg, When Offline Stores Reduce Online Returns

2021011-MARK: Hirche, C.F., M.J. Gijsenberg, and T.H.A. Bijmolt, Promoting Product Returns: Effects of Price Reductions on Customer Return Behavior

2021012-MARK: Hirche, C.F., M.J. Gijsenberg, and T.H.A. Bijmolt, Asking Less, Getting More? The Influence of Fixed-Fee and Threshold-Based Free Shipping on Online Orders and Returns

2021013-I&O: Sorgner, A., and M. Wyrwich, Calling Baumol: What Telephones Can Tell Us about the Allocation of Entrepreneurial Talent in the Face of Radical Institutional Changes

2021014-I&O: Slavtchev, V., and M. Wywich, TV and Entrepreneurship

2021015-EEF: Kate, F. ten, M.J. Klasing, and P. Milionis, Diversity, Identity and Tax Morale

2021016-EEF: Bergemann, A., and R.T. Riphahn, Maternal Employment Effects of Paid Parental Leave

2021017-GEM: Abolhassani, M., Productivity Spillovers of Multinational Enterprises through Worker Mobility: New Evidence for the Netherlands

2021018-GEM: Abolhassani, M., Productivity Spillovers of Superior Firms through Worker Mobility

2022001-GEM: Oosterhaven, J., A Price Reinterpretation of the Leontief Quantity Model

2022002-EEF: Ghaemi, S, X. Li, and M. Mulder, Economic Value of Flexibility Provided by Power to gas Conversion Systems in Low-voltage Distribution Grids

2022003-OB: Meer, P.H. van der, Are All Self-employed Happy?

2022004-EEF: Perey, P., and M. Mulder, International Competitiveness of Low-carbon Hydrogen Supply to the North-west European Market

2022005-OPERA: Kasper, A., M. Land, and R. Teunter, University of Groningen, Faculty of Economics and Business, Department of Operations.

2022006-I&O: Fritsch, M., and M. Wyrwich, Entrepreneurship in the Long-run: Empirical Evidence and Historical Mechanisms.

2022007-EEF: Treurniet, M., and R. Lensink, Belief-based Poverty Traps and the Effects of Material and Psychological Development Interventions.

2022008-EEF: Kwaak, Christiaan van der, Monetary Financing Does Not Produce Miraculous Fiscal Multipliers

2022009-EEF: Beesten, E.R. van, and D. Hulshof, Transmission Capacity Reduction in International Power Systems: Economic Incentives and Welfare Effects

2022010-OPERA: Romeijnders, W., N.D. van Foreest and J. Wijngaard, On Proportionally Fair Solutions for the Divorced-Parents Problem

2022011-I&O: Hipp, A., M. Fritsch, M. Greve, J. Günther, M. Lange, C. Liutik, B. Pfeifer, M. Shkolnykova, and M. Wyrwich, Comprehensive Patent Data of the German Democratic Republic 1949-1990—Technical Report and Dataset Overview

2022012-EEF: Bianchi-Vimercati, R., G. Lecce, and M. Magnaricotte, Persistent Specialization and Growth: The Italian Land Reform

2023001-I&O: Fritsch, M., M. Greve, and M. Wyrwich, Shades of a Socialist Legacy? Innovation Activity in East and West Germany 1877-2014

2023002-EEF: Bahcivan, H., L. Dam and H. Gonenc, New Avenues in Expected Returns: Investor Overreaction and Overnight Price Jumps in US Stock Markets

2023003-I&O: Wyrwich, M. and M. Fritsch, How does Regional Entrepreneurship Transfer over Time? The Role of Household Size and Economic Success

2023004-OPERA: Seepma, A., and M. Spring, The Roles of Technology and Co-location in Overcoming Divergent Logics in Professional Supply Chains: the Case of Criminal Justice

2023005-EEF: Dalo, A., R. Mees, B. Scholtens, and H. Zhao, Corporate Sustainability, Cost of Equity, and Credit Ratings

2023006-EEF: Kwaak, C. van der, and S. van Wijnbergen, Financial Fragility and the Fiscal Multiplier

2024001-EEF: Herrmann, O., A partisan duty to vote?

2024002-EEF: Soetevent, A.R., and G.J. Romensen, The Short and Long Term Effects of In-Person Performance Feedback: Evidence from a Large Bus Driver Coaching Program www.rug.nl/feb