INSEAD The Business School for the World®

Working Paper
2019/52/TOM

Nudging Drivers to Safety: Evidence from a Field Experiment

Vivek Choudhary
INSEAD, vivek.choudhary@insead.edu

Masha Shunko
Foster School of Business, mshunko@uw.edu

Serguei Netessine
The Wharton School, netessin@wharton.upenn.edu

Seongjoon Koo
J.D. Power, joon.koo@jdpa.com

Driving is an integral component of many operational systems and any small improvement in driving quality can have a significant effect on accidents, traffic, pollution, and the economy in general. However, making improvements is challenging given the complexity and multidimensionality of driving as a task. We use telematics technology (i.e., real-time sensor data in a mobile device such as accelerometer and gyroscope) to measure driving performance as well as to deliver nudges to the drivers via notifications. Leveraging a smartphone application launched by our industry partners, we sent three types of performance nudges to drivers, indicating how they performed on the current trip with respect to their personal best, personal average, and latest driving performance. We find that personal best and personal average nudges improve driving performance, on average, by 18.17% and 18.71% standard deviations of the performance scores calculated by the application. This improvement translates into an increase in the inter-accident time by nearly 1.8 years, while also improving driving performance consistency as measured by the coefficient of variation of the performance score. Using generalized random forests we show that high-performing drivers who are not frequent feedback seekers benefit the most from personal best nudges, while low-performing drivers who are also frequent feedback seekers benefit the most from the personal average nudges. Using these findings, we construct personalized nudges that outperform both of these nudges.

Key words: Nudges, Empirical Operations Management, Behavioral Operations Management, Field Experiments.

Nudging Drivers to Safety: Evidence from a Field Experiment

Abstract

Driving is an integral component of many operational systems and any small improvement in driving quality can have a significant effect on accidents, traffic, pollution, and the economy in general. However, making improvements is challenging given the complexity and multidimensionality of driving as a task.

We use telematics technology (i.e., real-time sensor data in a mobile device such as accelerometer and gyroscope) to measure driving performance as well as to deliver nudges to the drivers via notifications. Leveraging a smartphone application launched by our industry partners, we sent three types of performance nudges to drivers, indicating how they performed on the current trip with respect to their personal best, personal average, and latest driving performance.

We find that *personal best* and *personal average* nudges improve driving performance, on average, by 18.17% and 18.71% standard deviations of the performance scores calculated by the application. This improvement translates into an increase in the inter-accident time by nearly 1.8 years, while also improving driving performance consistency as measured by the coefficient of variation of the performance score. Using generalized random forests we show that high-performing drivers who are not frequent feedback seekers benefit the most from *personal best* nudges, while low-performing drivers who are also frequent feedback seekers benefit the most from the *personal average* nudges. Using these findings, we construct personalized nudges that outperform both of these nudges.

Keywords: Nudges, Empirical Operations Management, Behavioral Operations Management, Field Experiments.

1. Introduction

Driving is a key component of many operational systems and it makes a significant contribution to the economy. Nearly 1.3 billion on-road motor vehicles move people and goods worldwide. At the same time, WHO estimated that these vehicles are associated with 1.35 million deaths annually due to accidents (WHO 2018), 11% of which are caused by behavioral issues (such as distracted driving) and are therefore preventable. The advent of the gig economy (e.g., Uber), crowdsourced deliveries (e.g., InstaCart), and ridesharing (e.g., BlaBlaCar) makes driving a core value of these businesses but also leads to higher congestion and vehicle utilization (Cramer and Krueger 2016). Furthermore, beyond safety, improvement in driving can help reduce pollution, vehicle wear and tear, and congestion, thus, massively impacting supply chains and the economy in general.

Of course, driving is a complex multidimensional task and it is not easy to motivate better driving. Consequently, governments impose regulations (e.g., speed limits) and punishments (e.g., fines) for violations of traffic rules. Industries that rely on driving are increasingly implementing financial and non-financial interventions to improve driving performance. For instance, automotive insurance

companies use financial incentives (insurance discounts) to reduce accident claims. Further, fleet companies use monitoring through GPS-enabled devices as well as financial incentives to promote better driving. These financial approaches are expensive and therefore may not be sustainable in the long term (FleetAnswers 2018), and effectiveness of rules and regulations is limited as accident statistics indicate (WHO 2018). In this paper, we attempt to answer the question: "Can we motivate better driving using simple (and free) nudges?"

Our study is motivated in part by the rich work done in the area of behavioral economics related to *nudges*. Thaler and Sunstein (2008, p6) define a nudge as "… any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting the fruit at eye level counts as a nudge. Banning junk food does not". There are many examples of nudges (e.g., setting the default investment percentage in a pension fund to be high) and there have been many successful applications of nudges in several domains such as finance (Anderson and Robinson 2017, Cronqvist and Thaler 2004), utilities (Allcott 2011, Charlier and Guerassimoff 2018) and health (Bhattacharya et al. 2017), among others. Most nudges studied in the literature are designed to promote unidirectional responses such as eat healthier, save more, or reduce electricity consumption. In contrast, driving is a complex activity and has many dimensions such as braking, acceleration, and speeding. Further, these behaviors are correlated with each other, e.g., speeding is correlated with braking and acceleration, and moreover drivers tend to inherently overestimate their driving abilities (Roy and Liersch 2013). As a result, giving drivers simple feedback on their driving (Choudhary et al. 2018) surprisingly leads to further deterioration in performance, on average. Therefore, improving driving appears to be inherently challenging.

In this paper, we implement a novel intervention to study driving performance in a field experiment. Specifically, we nudge drivers into improving their driving performance by comparing their most current driving score with their personal best, personal average, or last performance. There are precedents of using nudges by the government to improve driving behavior. Recently, the Federal Highway Administration, which is a division of the United States Department of Transportation, granted a team that includes researchers from the University of Pennsylvania $1.84 million to study and implement nudge-like interventions to curb distracted driving (PennMedNews 2018). There are other implementations in practice such as life-size card board cutouts of policemen placed on the roads, displays with real-time speed, white lines with narrowing gap to provide illusion of higher speed etc. (BBC 2014). To our knowledge, there are currently no rigorous academic studies of the effectiveness of these interventions.

To understand the effectiveness of different types of nudges on driving performance, we leverage a telematics smartphone application (app) to deliver nudges to 1069 drivers in a randomized field experiment in India. The app uses real-time data from the embedded sensors in a mobile device (such as an accelerometer, GPS, and gyroscope) to measure driving performance. Specifically, we consider

three types of interventions – *personal best, personal average,* and *latest score* performance nudges by placing drivers into three treatment groups and one control group.

We find that *personal best* and *personal average* nudges improve driving performance by 18.17% and 18.71% respectively, in terms of the standard deviation of their trip performance. The absolute increase in driving performance by the two nudges on an average prolongs the inter-accident time by nearly 1.8 years as estimated by our industry partner. Additionally, we find that *personal best* and *personal average* nudges result in improved driving consistency as measured by the coefficient of variation of trip performance. We do not find any such effect for *last score nudge.*

Beyond this average treatment effect, we study the heterogeneous treatment effect of nudges using machine learning approach for causal inference, i.e., generalized random forest (Athey et al. 2019). We show that both personal best and personal average nudges benefit low-performing drivers more than high-performing drivers. Moreover, high feedback-seeking drivers benefit from personal average nudges whereas all types of feedback seekers benefit from personal best nudges, on average. We further propose personalization of nudges based on the drivers' profile (average performance and feedback-seeking behavior) which can increase the treatment effect by nearly 11%.

This paper makes several contributions: 1) using field experiments, we causally estimate the effect of different types of performance nudges on driving behavior, 2) we contribute to the operations management literature by providing evidence that inexpensive interventions can work to improve complex operational tasks such as driving, 3) we apply a novel methodology (generalized random forest) to study heterogeneous treatment effects, and 4) we propose a novel personalized nudging intervention and estimate its effect.

2. Related Literature

Our paper is closely related to the streams of literature in psychology and economics that study effect of nudges. Extensive work implementing nudge interventions has found them to be effective in improving performance across various domains such as finance, utilities, transportation, and health. We cannot possibly cover all the work done in this area, so we refer our readers to Thaler and Sunstein (2008) for an overview. Broadly speaking, studies in this area demonstrate that inexpensive, easy to implement, and simple interventions can be very effective in improving performance. For example, using a flagship saving-commitment program called *save more tomorrow,* Thaler and Benartzi (2004) report that using default nudges can result in significant improvement in saving rates (an increase from 3.5% to 13.6%). A recent paper by Kagan et al. (2018) shows that nudging can be effective in improving performance in a task where a team performs both design and execution (i.e., a complex operational task). In a lab experiment, they show that nudges improve performance by reducing the delays in transitioning from the design to execution phase. Similar to their task at hand, we motivate drivers in a complex process (i.e., driving) through nudges. Our study is different as we do not study teams and we use a field experiment.

Similar to any behavioral intervention, it is important to highlight that nudges should be crafted carefully to have the intended effect. For instance, Schultz et al. (2007) show that nudging users to consume less electricity through social norms may backfire for low-consumption users, who increased their consumption after receiving the nudge. Similarly, healthy labeling of food may have unintended consequences of excess consumption (Marteau et al. 2011). Sunstein (2017) elaborates further on when nudges can be ineffective.

Within the transportation domain specifically, some nudges have been found to be effective in improving driving behavior. For instance, in a field study on Kenyan drivers, Habyarimana and Jack (2015) find that nudging bus passengers to voice their concerns through placing complaint stickers on the bus lowers accident rates. This result was driven by the empowerment of passengers, whereby they could call a phone number to lodge a complaint against a reckless driver. However, this application is limited to fleets of trucks or buses. Further, several studies in this domain investigate how people make environmentally friendly choices through lab experiments. For example, Camilleri and Larrick (2014) use fuel economy labels to promote the choice of vehicles with lower fuel consumption. Similarly, Gaker et al. (2010) report that providing greenhouse gas emission information can lead drivers to select the energy efficient route. Unlike these lab studies, we conduct our experiment in the field and we use technology-enabled nudges that are related to the immediate past performance rather than to distant past performance.

We provide three types of nudges to the drivers using reference points anchored to own past performance, i.e., personal best, personal average, and latest performance. All reference points are tied to the trip performance or trip score -- an aggregate measure that encompasses multiple dimensions of driving performance. Therefore, our study is closely related to the papers that analyze the effect of information tied to a performance reference point that is shared with users in an attempt to improve performance.

Our first nudge references the personal best trip score achieved by a driver (*personal best nudge).* Personal best as a reference point has been found to be effective in improving performance in settings such as sports and education. Studying chess (also a complex task), Anderson and Green (2018) found that players exert more effort to set new personal best performance records. The literature on goal setting suggests that specific and difficult goals induce effort (Locke and Latham 2002) and therefore, personal best goals have been shown to improve performance. For example, Burns et al. (2018) study Australian adolescent students and find that personal best goals can improve engagement during the course of secondary school. Furthering the understanding of personal best goals, Martin (2006) suggests that personal best encompasses competitively self-referenced (i.e., people *compete* with *own* peak performance) and self-improvement (i.e., people *focus* on *own* performance) goal. This study of students shows that personal best goals lead to higher engagement on various dimensions. Further, it has been found that in sports, personal best score is the key indicator of 24-hour endurance run success rather than the actual physical attributes of the runner (Knechtle et al. 2009). As our first nudge reminds users

about their personal best performance (peak) and compares it with their current (latest) performance, the personal best nudge is also related to the literature studying *peak-end* anchoring (Fredrickson and Kahneman 1993). This literature shows that people tend to make decisions based on the recollection of the most intense and the most recent experience, where the most intense experience may be either positive or negative. Our nudge highlights the most intense *positive* experience potentially impacting this bias along with the most recent experience – we then test how this bias impacts the driving performance.

Our second nudge uses drivers' average trip scores for all past trips as a reference, which we call the *personal average nudge.* Average performance information has been used in the literature to provide assessment of users and induce performance, but in the form of feedback rather than as a nudge. For instance, in an aiming task, Yao et al. (1994) report that providing average performance to the subjects improves learning and retention compared to no feedback. Further, in a setup to promote eco-driving, Stillwater and Kurani (2013) report that providing information on average performance on miles per gallon leads to goal-setting behavior as well as goal achievement. In the same study, they found that providing immediate performance leads to experimentation in driving behavior, which is closely related to our third type of nudge, *last score nudge*. However, we are not aware of any studies that test personal average nudges specifically.

Our third nudge intervention is motivated by similar studies providing last score performance in other settings. Many studies in the operations management literature appraise their subjects with performance in the last decision. Such performance feedback information differs from nudges, as feedback does not call for a specific action such as "aim to beat your personal best score" as we do. For instance, in newsvendor games (Bolton and Katok 2008) players are provided with their latest profits after each game round. Similarly, experiments studying the bullwhip effect (Croson and Donohue 2006) provide information to the players on inventory after each round. Further, in a newsvendor lab experiment Kremer et al. (2011) provide on-screen feedback on past performance after each round to study forecasting biases and performance, which is related to our last score nudge intervention. Although last score provided to the subjects is an important feedback, these papers study inherent decision biases in the presence of such information whereas we study the effect of such information as well as of the direction to act, articulated in the form of nudges. Our last intervention is also closely related to the contemporary telematics feedback systems implemented in practice. For example, in the telematics application used by many automotive insurance firms, users are provided with their latest trip score after each trip completion. However, it is important to note that simple feedback on driving does not seem to improve drivers' behavior and, in fact, it does the opposite (Choudhary et al. 2018). Moreover, it has been shown that frequent feedback through real-time information systems may deteriorate performance (Lurie and Swaminathan 2009). Thus, there is a need to devise different (stronger) interventions such as nudges. As a result, we reference our third treatment to the latest trip score achieved by a driver and we call it a *last score nudge.*

We contribute to the OM literature that employs various behavioral interventions such as feedback to improve performance. For instance, Song et al. (2018) study social comparisons (public relative performance) to improve physicians' productivity in a field setup. In a lab experiment, Kim et al. (2018) show that removing information hurdle, i.e., making information salient, can improve hospital bed utilization by correcting physicians' belief regarding current bed utilization. Similarly, Bendoly (2013) studies the effect of real-time feedback on performance (acceptance of a bid) in a hospitality set up (hotel) in a lab experiment. Given that people show a tendency of *system neglect* (Massey and Wu 2005) by paying too much attention to signals as well as inadequately update their beliefs about themselves (Markus et al. 2014), our nudges provide an alternative and inexpensive intervention to reduce such biases and subsequently improve performance.

Our study is realted to the literature that studies different types of inexpensive interventions (such as priming in Balafoutas et al. (2018)) to improve individual performance. Broadly, our paper pursues the 'improve behavior' category, which behavioral operations management scholars outline as a key research goal (Donohue et al. 2019). Methodologically, we employ a field experiment, which has been recognized as an important method to study behavioral operations management problems (Ibanez and Staats 2018). We specifically focus on drivers, similar to papers that tested other types of incentives in the field to improve driving and safety. For instance, Bolderdijk et al. (2011) report that young drivers' speeding behavior improves significantly using a pay-as-you-drive insurance. Similarly, in a field experiment Chen et al. (2017) estimated that sending social comparison text messages can reduce traffic rule violations by 5 to 6%. Yang and Long Lim (2018) use discounts (in ticket prices) to reduce congestion in a subway. Our study aids our understanding of human behavior when using the telematics applications (Soleymanian et al. 2019) and further develops understanding of nudges for complex tasks. Our paper adds to the recent OM literature that addresses complex operational problems such as improving organ donation rates (Tayur 2017, Tayur et al. 2019) and innovation performance (Kagan et al. 2018) through nudges.

3. Theory and Design of Interventions

A recent scoping survey (Szaszi et al. 2018) highlights various moderating roles of nudges in improving performance, noting specifically two mechanisms – 1) making information salient (e.g., making the information easily available) and 2) reminding users about the purpose (e.g., visit a gym more often). Thus, when nudging, we provide users with their latest trip performance and we compare it to the driver's reference performance (i.e., personal best, average, or last score score). If a driver beats her reference point on the latest trip, we provide information on the reference point and nudge the driver to keep progressing. If a driver performs below his/her reference point, we nudge the driver to beat the reference by providing the reference point information as well. Therefore, we make the reference point salient as well as remind the users about the purpose (improve driving) through these nudges.

Making performance salient has several advantages that are studied in the goal-setting and feedback

literature. Importantly, saliency promotes anchoring through a reference point (Anderson and Green 2018), which promotes goal setting. It is important to highlight that our reference points are related to one's past performance. Goals based on own past performance are deemed achievable through increased self-efficacy (Bandura 2010). Therefore, we postulate that self-referenced nudges will promote goal setting and induce better performance.

Moreover, through these nudges we also provide feedback and therefore help calibrate beliefs of drivers of their own ability (Du et al. 2012). Calibration of such beliefs is important because driving is unique in that drivers tend to overestimate their driving capabilities (Roy and Liersch 2013). Such overconfidence in abilities may lead to underestimation of risk associated with poor driving (Helweg-Larsen and Shepperd 2001) as well as to exerting less effort to improve their driving behavior. Nudges could help in calibrating self-beliefs and in motivating drivers to perform better. Although all three nudges are designed to improve performance, they are distinct in their design.

Our first nudge (*personal best nudge*) reminds users about their peak performance and thus, provides an aspirational goal (Martin 2006). Personal best is non-decreasing over time and in most cases is higher than or equal to the current performance (unlike two other nudges that we test), and we hypothesize that it should provide the strongest motivation to improve for the drivers.

Our second nudge (*personal average nudge*) is easier to achieve since it is not non-decreasing. Similar to our argument about personal best nudge, personal average nudge should provide motivation to improve by reminding users about their average performance through attainable goal setting. However, we expect that its strength is somewhat lower than for the personal best nudge since personal average is clearly at or below the personal best.

Our third nudge (*last score nudge*) refers to recent performance. By providing a last score nudge, we equip drivers with information about their scores on their two recent trips and a comparison between them. In a similar setting but without nudges, Choudhary et al. (2018) show that drivers observing an increase in their recent score tend to perform worse on the next trip while a large decrease in score leads to better performance. In our case, we not only provide the two latest scores but also, in addition, we nudge drivers to beat their last score. When drivers beat their last score, we nudge them to keep progressing. Therefore, providing users with their latest performance and a nudge can help in two ways to have a positive effect on performance. First, if drivers perform worse on the trip than on the previous one, they will tend to exert more effort to match their past performance. Second, if drivers have done better on the last trip then they might exert enough effort to maintain their performance due to the nudge (to keep progressing). In both cases, drivers seeing last score nudge should perform better than the control group with no nudges. That said, last score nudge has two relative disadvantages: first, it does not account for any history beyond last trip, while the other two types of nudges do, and second, mathematically, last score nudge has higher variance than the other two types of nudges. Consistency in nudges (i.e., lower variation) should lead to better calibration of belief as the users receive consistent signal on their performance. On the other hand, high variation in feedback can lead to ignoring feedback

that is negative (DeJoy 1992, Grossman and Owens 2012). Hence, we expect that the last score nudge will have the lowest impact on performance relative to the personal best and personal average nudges.

To summarize, ex-ante, we predict that the personal average and personal best nudges should have a larger effect on performance than the last score nudge, and perhaps personal best should be preferred over personal average.

4. The Field Experiment

We investigate the effect of *performance nudges* on driving performance by conducting a field experiment with drivers in India.

Data Collection: On 22 May 2018, Raxel Telematics (Raxel) in collaboration with J.D.Power (JDPower) launched a smartphone application (app) DrivePower. The main objective was to test usage of a telematics app in the Indian market. The drivers were recruited mostly through social media (Facebook) and SMS campaigns although the app also provides an option to use referrals for sign-up. Once drivers downloaded the app, they would receive immediate feedback on their driving performance. In the first few trips, the app creates unique driving signatures using state-of-the-art machine-learning algorithm to identify whether a driver is driving or if s/he is a passenger in the car. The app is also capable of identifying different modes of transport, e.g., if users are on a bus, train, or bike. These algorithms are routinely used in automotive insurance industry. The app also provided incentives to participate in product surveys (such as evaluating car seats and audio systems) and an option to participate in competitions and earn cash (e.g., invite 20 users and earn INR 50). Each driver was required to drive a minimum of 500km and answer the survey questions for JDPower to earn monetary incentives. There were no incentives related to the driving performance. When users sign-up, the app explains different components of driving, how the application will provide feedback on a trip basis, the surveys, different screens, and other terms and conditions for the app. The DrivePower website (http://mydrivepower.com/) was created to share the details of the program as well as to explain the various measures of driving performance. To protect the anonymity of the users, Raxel provided us with the UserId (a unique number created for each driver) and removed all personal identifiers: this anonymized data was used for our analysis.

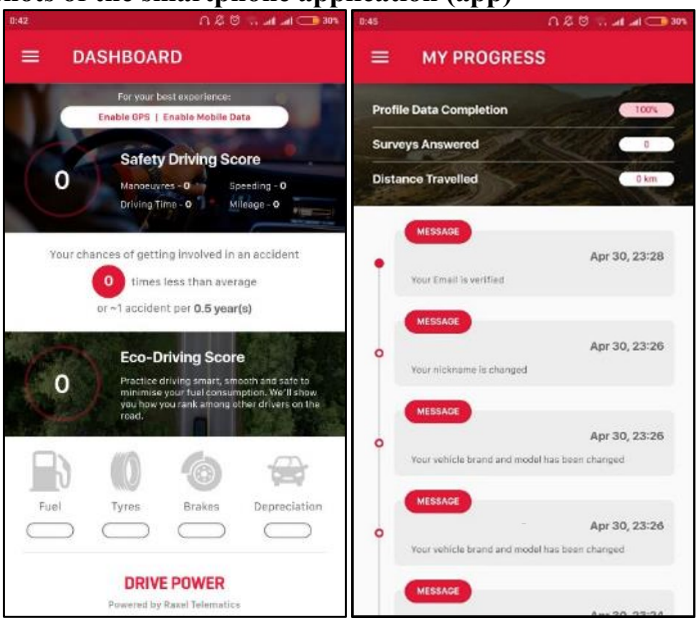
Smartphone Application (app): We leveraged the app launched by Raxel and JDPower on both iOS and Android by providing different nudges to the registered users to study the effect of performance nudges on driving performance. The app consists of a main dashboard and several other screens (sample screenshots provided in Figure 1). The first screenshot is the landing screen for the drivers, which is called *dashboard* and has two parts. The upper half displays the *Safety Driving Score,* an aggregated to-date score on a scale of 0 to 100. After each trip, users are notified about their latest trip score via a push notification. Raxel uses a proprietary algorithm to calculate trip score based on various components of driving; subsequently, individual trip scores are aggregated to create the Safety Driving Score. Note that the dashboard also shows the to-date scores across four dimensions. *Maneuver* score denotes the

combined maneuver behavior consisting of harsh braking and sharp accelerations. *Speeding* denotes the speeding behavior in driving, which depends on the duration and distance covered while speeding as well as the difference between actual speed and the speed limit. *Driving Time* denotes the score based on times of the day trips are taken with rush hours and night hours penalized as they are associated with higher probability of an accident. *Mileage* denotes the distance travelled by a driver with longer distances travelled penalized because these are associated with higher probability of an accident. All these scores are provided on a scale from 0 to 100, where 100 denotes the best performance.

In the lower part of the screen, the driver sees the *eco score,* which is a safety driving score transformed for easier understanding of short-term metrics for drivers such as fuel consumption and maintenance costs. The eco score is similar to the indicators on modern car dashboards that show the optimal speeds for higher fuel efficiency. Our main outcome variable is trip score (i.e., safety driving score for each trip), which we refer to as *score* in the rest of the document. Note that only safety score is sent as a notification and not the eco score.

The second screenshot shows *my progress* screen. All trip scores are recorded here, which are used to calculate safety driving score. A user can click on a particular trip and open the screen to check the respective score *in that* trip across the four dimensions mentioned above. This screen also displays all the nudges sent to the users, which were sent via push notifications. To make sure that *my progress* screen is not overpopulated with messages, only specific messages are available here (e.g., trip score, nudges) and not all notifications that were sent to the driver. One specific attribute of our app is that we can track which screens have been visited. Therefore, we can understand how/when users interact with the app over time.

Figure 1: Screenshots of the smartphone application (app)



Launch process: After collecting baseline data for nearly three months, on 25 July 2018 we launched

our experiment. We randomly assigned all the drivers in the program into four groups – three treatment groups and a control group. The three treatment groups received three types of performance nudges – 1) Personal best nudges (*PBnudge)*, 2) Average performance nudges (*AVnudge*), and 3) Last score nudges (*LSnudge*). All nudges were sent as notifications to the users and were available in *my progress* screen for review at a later time. To focus on the pure effect of nudges, we did not provide drivers with any incentives based on performance. We collected driving data for these users until 18 October 2018, at which time Raxel sent messages to everyone to redeem monetary rewards (for participating in the program – reward did not depend on performance) and drivers started exiting the platform.

Drivers' details: Our final dataset consists of 1069 drivers, of which 529 drivers (our main subjects) took trips both before and after the launch of the experiment. These 529 drivers (both in treatment groups and control group) clocked 105,101 trips in total. We report the distribution of these drivers and their number of trips across treatments in Table 1. This pool of drivers excludes seven drivers who left the program after we randomized and stopped taking trips within a week from the start of the treatment (1 from *PBnudge,* 2 from *AVnudge*, 2 from *LSnudge,* and 3 from *control).* We did not find any significant difference for these drivers in terms of their observed characteristics such as performance and distance travelled. We use the remaining 540 drivers for a robustness check (refer to section 7.4 for details).

We have participants from 536 out of the 701 districts in India. However, nearly one third of the trips were made in the following major locations -- Bengaluru, Pune, Ahmedabad, Gurugram, Mumbai (including suburban), Thane, and Delhi.

Table 1: Number of drivers and trips across groups

	Control	*Treatment*			*Total*
		Personal best nudge (PBnudge)	*Average nudge (AVnudge)*	*Last score nudge (LSnudge)*	
#drivers	149	109	140	131	529
#trips	25,567	25,777	26,442	27,315	105,101

Nudges: We programmed backend to send nudges using *push notifications* after the trip has ended. Note that these nudges are significantly different form the trip score notification (e.g., "your trip score was 80"). As described earlier, these nudges also appear in *my progress* screen for drivers to review later. The objective was to increase visibility of these nudges. Nudges were sent every three days at the same time (10 am IST)[1]. We call these *nudge-days*. We chose the same time and frequency that Raxel used to send neutral messages such as "Thank you for joining DrivePower! Keep driving." prior to the experiment launch. We only altered the messages for the treatment groups to identify the effect of nudges.

We computed the reference points for each nudge type as follows. Consider a driver *i* who took *j*

[1] For safety, nudges are sent only when drivers are not driving. In rare cases when a nudge is sent while driving, it will be delayed until after the trip is over.

trips since the beginning of the program until 10am on a *nudge-day.* The reference point for driver *i* then is given by the following formulae where $rating_{ij}$ denotes trip score of driver *i* in trip *j*:

$$reference_{ij} = \begin{cases} \text{mean}(rating_{i1}, \ldots., rating_{ij-1}) & if\ i \in nudgeAV; \\ \max(rating_{i1}, \ldots., rating_{ij-1}) & if\ i \in nudgePB; \\ rating_{ij-1} & if\ i \in nudgeLS. \end{cases}$$

Driver *i* then would receive a nudge as stated in Table 2. Note that both types of messages are nudges as they provide a comparison to the reference and urge users to improve their performance either by progressing with their current achievement (of beating the reference point) or aiming to beat their reference point in future. The control group kept on receiving the same neutral message from Raxel to ensure that the effect that we observe is only from the changes in the treatment messages.

Variables: We now describe the variables that we use in this study.

score -- the score obtained by a driver on the last trip and delivered to them via the app within a few seconds of finishing the trip. There are six possible values for this variable (0, 20, 40, 60, 80, and 100) calculated using a proprietary algorithm. The algorithm takes into account the number of harsh breakings, sharp accelerations, speeding instances (duration, distance, and extent of over speed), time of the day, distance traveled, as well as driving duration and assigns a discrete score. The higher the score the better the driving performance.

todatescore -- the aggregated score of all the trips completed in the past (it appears as *safety driving score* on the main dashboard, top panel). This aggregation is also done based on a proprietary algorithm that accounts for distance and duration when weighing each of the individual trip scores. Automotive insurance firms focus on this score, which is used as a predictor of possible claims in the future.

Table 2: Description of nudges

Group	If $rating_{ij} > reference_{ij}$
AVnudge	Your safety driving score was $rating_{ij}$ recently. You beat your personal average score of $reference_{ij}$, keep progressing!
LSnudge	Your safety driving score was $rating_{ij}$ recently. You beat your last trip score of $reference_{ij}$, keep progressing!
PBnudge	Your safety driving score was $rating_{ij}$ recently. You beat your personal best score of $reference_{ij}$, keep progressing!
	If $rating_{ij} \leq reference_{ij}$
AVnudge	Your safety driving score was $rating_{ij}$ recently. Aim to beat your personal average score of $reference_{ij}$, keep progressing!
LSnudge	Your safety driving score was $rating_{ij}$ recently. Aim to beat your last trip score of $reference_{ij}$, keep progressing!
PBnudge	Your safety driving score was $rating_{ij}$ recently. Aim to beat your personal best score of $reference_{ij}$, keep progressing!

distance -- the last trip length in kilometers captured via GPS.

dayhours (nighthours, rushhours) -- the duration in minutes of the portion of the last trip travelled during day (night, rush) hours. Each day is broken into three mutually exclusive time intervals: day,

night, and rush; the cutoffs for the time intervals depend on the day of the week and holidays. Weekends and holidays do not have rush hours.

daystillexp -- number of days in the program before the start of the experiment. This is a proxy for the experience with the app that can affect score due to learning effect from the feedback provided via the app.

tripsbefore -- number of trips in the program before the beginning of the experiment. Similar to the *daystillexp* we use this variable to capture learning effect from the feedback provided via the app. This variable together with *daystillexp* captures driving frequency.

fb_bf_dashboard -- a binary variable capturing whether a user has opened the app and looked at the dashboard before a trip (=1) or not (=0).

Table 3: Balance table for all groups

	(1) *control*	(2) *AVnudge*	(3) *LSnudge*	(4) *PBnudge*	*t*-test Difference		
Variable					*(1)-(2)*	(1)-(3)	(1)-(4)
score	76.09	78.07	77.70	79.05	-1.98	-1.61	-2.96
	(1.43)	(0.93)	(1.29)	(1.08)			
todatescore	72.06	73.80	73.53	75.30	-1.74	-1.47	-3.24
	(1.70)	(1.30)	(1.47)	(1.48)			
distance	12.43	11.66	12.08	11.33	0.77	0.36	1.10
	(0.58)	(0.53)	(0.53)	(0.50)			
dayhours	17.29	16.59	18.31	18.19	0.69	-1.03	-0.90
	(0.64)	(0.59)	(0.90)	(0.78)			
nighthours	1.11	1.29	1.20	1.23	-0.18	-0.09	-0.12
	(0.15)	(0.18)	(0.18)	(0.21)			
rushhours	7.22	7.17	8.04	7.36	0.06	-0.82	-0.13
	(0.46)	(0.43)	(0.47)	(0.42)			
fb_bf_dashboard	0.14	0.17	0.17	0.14	-0.02	-0.03	0.01
	(0.02)	(0.02)	(0.02)	(0.02)			
daystillexp	50.40	48.83	50.87	51.34	1.57	-0.47	-0.94
	(1.53)	(1.55)	(1.61)	(1.46)			
tripsbefore	201.47	177.03	200.04	227.09	24.43	1.43	-25.63
	(16.72)	(11.96)	(16.35)	(19.44)			
N	12799	11768	12984	12771			

*Standard errors in parentheses, *** p<0.01, ** p<0.05, *p<0.1*

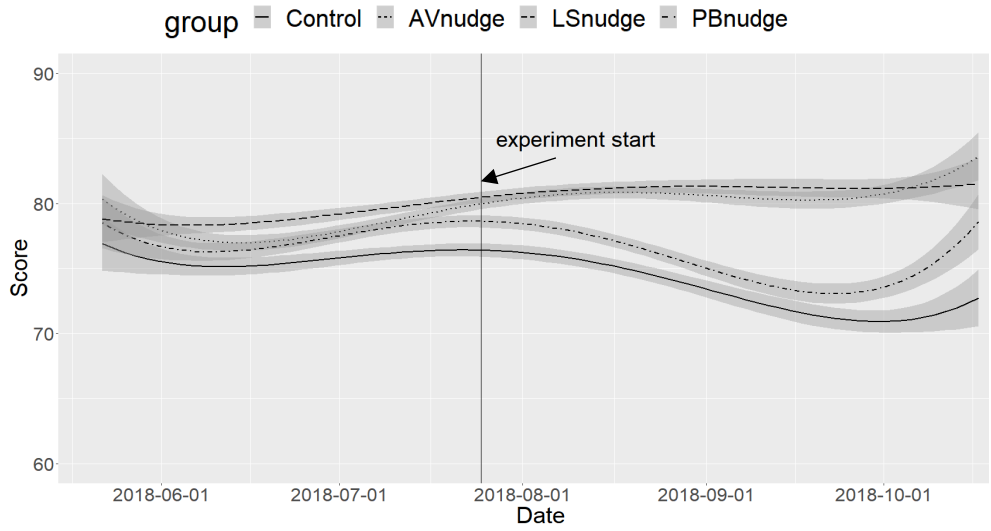
weekday -- a binary variable that captures whether a trip has been taken on a weekday (=1) or not (=0). We use this variable to capture the differences in traffic behavior between a weekday and a non-week day.

after -- a binary indicator to capture if a trip has been taken *after* the launch of the experiment.

Randomization and Variable Balance: We randomized users before the launch of our experiment using a four-faced dice with equal probability. To test the quality of our randomization, we check for balance in the observed variables and report the results in Table 3. We do not find any statistically significant differences in the pre-experiment variables. In the first four columns we report the group averages as well as standard errors (in parentheses) below each mean. The last three columns show the *t*-test results for the differences in means. We cluster the standard errors at the individual driver level to account for the possible correlation for the repeated observations for the same driver. It is important to

highlight that significant differences can be observed even after randomization for some parameters just by chance and controlling for those variables can address such issues.

Figure 2: Visual check for parallel trend between treatment and control



5. Results

First, we provide a model-free evidence of the effect of nudges on performance as measured by trip score. Further, we use difference-in-difference models to estimate the effect of nudges on trip performance. This analysis helps explain the *average* effect of nudges. Next, we recognize that driving encompasses negative externalities: a reduction in performance of any driver can lead to unsafe driving conditions for other drivers on the road. Hence, understanding of average effects may be insufficient: for instance, it may happen that we observe a positive effect of nudges because some drivers performed better while others did worse. Therefore, we finally investigate the heterogeneous effect of nudges on performance using generalized random forest.

Table 4: Before and after trip mean performance

	Variable	(1) pre-treatment	(2) post-treatment	t-test Difference (2)-(1)
Control	*score*	76.09	73.82	-2.28**
		(1.43)	(1.55)	
	N	12799	12768	
AVnudge	*score*	78.07	80.78	2.71***
		(0.93)	(0.75)	
	N	11739	14635	
LSnudge	*score*	77.70	76.18	-1.52
		(1.29)	(1.53)	
	N	12984	14331	
PBnudge	*score*	79.05	81.24	2.19**
		(1.08)	(0.92)	
	N	12771	13006	

*Standard errors in parentheses, *** p<0.01, ** p<0.05, *p<0.1*

5.1 Model-Free Evidence

In Figure 2, we plot the average trip score of the three treatment groups and the control group over time. We observe that all four groups are very close to each other in terms of their pre-treatment average scores. After the launch of the experiment, the control group has a downward trend, while two of the three groups, i.e., *AVnudge* and *PBnudge* exhibit improvement in average scores. However, group *LSnudge* does not show any improvement and almost follows the trend of the control group. Overall, the plots in Figure 2 suggest that not all nudges are effective in improving driving performance. In addition to the daily average scores, we also plot weekly averages for individuals and observe similar trend to the one reported in Figure 2.

Next, we study the pre- and post-experiment means for all four groups, and we test the changes statistically. As evident from Table 4, only two out of the three nudges (*AVnudge* and *PBnudge)* have positive effect on performance. On an average, when control group scores decrease by 2.28 points, *AVnudge* and *PBnudge* lead to 2.19 and 2.71 points improvement in the trip score respectively. *LSnudge* on average does not show statistically significant difference in pre- and post- means. We will now use a difference-in-difference model to study these effects.

5.2 Empirical Model

Given that we have pre- and post-experiment observations, we use difference-in-differences (DID) models to estimate the size of the treatment effect. Before we delve into the effect size estimation, we have to check the fundamental assumption of parallel trends (Terwiesch et al. 2019) between the control and the treated groups in the pre-treatment time period. It is apparent from visual inspection of Figure 2 that our three groups approximately show parallel trends prior to the start of the experiment. As a robustness test, we also estimate the effect size using the *Synthetic Control Method* (for details of the methodology refer to Abadie et al. (2015)), which relaxes the assumption of parallel trends for identification (refer to the robustness section for details). We estimate the following linear model to test the effect of our treatment:

$$score_{igjt} = \beta_0 + \beta_1 TreatmentDummy_{ig} + \beta_2\left(TreatmentDummy_{ig} \times after_t\right) + \beta_3 after_t + \beta_4 Controls_{igjt} + \beta_5 weekday_t + \epsilon_{igjt}, \quad (1)$$

where *i* denotes an individual, *g* denotes a group (control, *LSnudge, PBnudge, or AVnudge*), *j* denotes a trip, and *t* denotes date. We use group dummies to control for unobservable differences among groups. In addition, we control for all variables mentioned in Table 3 for a more conservative estimation of the treatment effect. As we use group-level dummies, these variables also help control for time-invariant heterogeneity among drivers such as the number of trips taken before the launch of the experiment. There are rush hours on weekdays, therefore, we include a binary weekday dummy to control for type of day (1 for weekday, 0 otherwise). This is important as users can travel on a weekday but not during a rush hour. Further, we define a binary variable *after* which is one if a trip has been taken after the treatment has started. Our coefficient of interest is β_2, which is the DID estimator that captures the effect

of our intervention. We report the results of our estimation in Table 5.

Table 5: Estimation of the effect of nudges on performance

	Dependent variable: score		
	(1)	(2)	(3)
distance	-0.62***	-0.62***	-0.06***
	(0.03)	(0.03)	(0.00)
dailyhours	0.13***	0.13***	0.01***
	(0.02)	(0.02)	(0.00)
nighthours	-0.24***	-0.24***	-0.03***
	(0.04)	(0.04)	(0.00)
rushhours	-0.12***	-0.12***	-0.01***
	(0.02)	(0.02)	(0.00)
fb_bf_dashboard	-1.87***	-1.81***	-0.17***
	(0.66)	(0.61)	(0.05)
daystillexp	-0.05*	-0.05*	-0.00**
	(0.03)	(0.03)	(0.00)
tripsbefore	-0.01	-0.01	-0.00
	(0.01)	(0.01)	(0.00)
weekday	-2.21***	-2.18***	-0.36***
	(0.28)	(0.28)	(0.03)
AVnudge		1.35	0.08
		(1.55)	(0.13)
LSnudge		1.42	0.09
		(1.77)	(0.14)
PBnudge		2.43	0.16
		(1.72)	(0.14)
AVnudge×after		4.44***	0.36***
		(1.26)	(0.10)
LSnudge×after		-0.00	0.05
		(1.39)	(0.11)
PBnudge×after		4.57***	0.35***
		(1.32)	(0.11)
after	-0.30	-2.66**	-0.23***
	(0.52)	(1.04)	(0.08)
Constant	89.87***	88.87***	
	(1.22)	(1.64)	
Estimation	*OLS*	*OLS*	*Ordered logit*

*Robust standard errors in parentheses, clustered at individual driver level *** p<0.01, ** p<0.05, *p<0.1, estimated with all controls.*

The first column is estimated with only control variables. Our results suggest that distance, night hours, and rush hours are negatively associated with performance. The coefficients of these variables are significant and negative (coefficients are -0.62, -0.24, and -0.12 respectively; all $p<0.01$). The day hour travel is associated with a positive effect of 0.13 (p<0.01). These coefficients are as anticipated. Further, we find that number of days till the experiment is associated with negative coefficient (-0.05), however the association is rather weak ($p<0.10$). We find no association of pre-treatment trips taken with the scores as suggested by the insignificant coefficient of *tripsbefore* variable. Furthermore,

weekdays are associated negatively with *score,* which may be due to the fact that the traffic is higher on a weekday compared to an average weekend. Lastly, the variable *fb_bf_dashboard* is worth highlighting here. The coefficient of the variable is -1.87 ($p<0.01$) implying that feedback taken before a trip negatively impacts performance by nearly 2 points. This result replicates the previously documented negative average effect of feedback on performance (Choudhary et al. 2018).

Our second model in Table 5 has been estimated with an OLS model (equation 1 above), using dummy variables for three treatment groups. Our results suggest that, compared to control group, there is no effect of being in one of the three groups on average (coefficients of *AVnudge, PBnudge, and LSnudge* are all insignificant). The coefficient of the variable *after* is negative and significant (-2.66, $p<0.05$), which signifies that, on average, there is a decline in performance of the drivers after the launch of the experiment.

Now, we focus on interpreting the interaction terms between treatment dummies and the variable *after,* which are our DID estimators. We find that, out of the three interaction terms, two are significant and positive. The DID estimators, i.e., coefficient of *AVnudge×after* is 4.44 ($p<0.01$) and *PBnudge×after* is 4.57 ($p<0.01$), meaning that, on average, the two nudges have a positive impact on performance. On average, these two treatment groups perform nearly 4.5 points better than the control group. In terms of the standard deviation of pre-treatment *score* (SD = 24.43) for the entire population (the standard deviations of scores are very similar between groups), the effect sizes suggest that, on average, the *personal average* and *personal best nudges* improve the performance by 18.17% (=4.44/24.43%) and 18.71% (=4.57/24.43%) standard deviation points respectively in terms of the pre-treatment population standard deviation. We do not find any effect of *LSnudge* on performance as evident from the insignificant coefficient of *LSnudge×after*. In summary, our results show that nudges work, however, not all of them.

With the help of our collaborating firm Raxel, which creates telematics application for the insurance firms, we can approximately determine the economic effect of our results. The performance parameter *score* has been found to be correlated with the probability of accidents and claims experienced by the insurance firms. We can, therefore, use score to understand the economic effect of improvement counted in number of years until the next accident for a given driver. For an average driver in our data set with a mean score of ~78, one score point increase in performance will increase the time until the next accident by 0.4 years, all else equal (i.e., all control variables remain the same). Therefore, our results indicate that performance nudges (*AVnudge* and *PBnudge)* that improve the scores by nearly 4.5 points can increase the inter-accident time by nearly 1.8 years or 21 months. This is a significant improvement from an individual, social, and business perspective.

Given that we have six categories in the outcome variable (0 to 100 in an interval of 20), to test the robustness, we compare the estimation results using an ordered logistics model. As reported in column 3, we find that our results are consistent. Due to the consistency of the results and ease of interpretation of OLS results, we will henceforth use a linear model for analysis (Wooldridge 2010, Yang and Long

Lim 2018). Finally, we also estimate our results with individual fixed effects to control for individual level time-invariant parameters. Our results are consistent with fixed effect specification as well (please refer to Table 9).

Consistency of driving – In addition to the effect of nudges on average trip performance, we are interested in finding whether nudges have an effect on variation in performance, i.e., do nudges affect driving consistency. To test this, we calculate pre- and post-experiment coefficient of variation of the score (where COV = *StDev(trip score)/Mean(trip score)*) as a measure of consistency for each driver. We then estimate the effect of different nudges on COV using OLS. Results in Table 6 suggest that, after the intervention, both *AVnudge* and *PBnudge* reduce the coefficient of variation relative to the control group. The coefficients for *AVnudge*×*after* and *PBnudge*×*after* are both -0.07 (*p<0.01*). The pre-experiment mean and standard deviation for COV are 0.29 and 0.11 respectively. Therefore, there is a 24% (=0.07/0.29%) reduction in COV due to the two nudges.

Table 6: Effect of nudges on coefficient of variation

	Dependent variable: COV
AVnudge	-0.02
	(0.02)
LSnudge	-0.02
	(0.02)
PBnudge	-0.04*
	(0.02)
AVnudgeXafter	-0.07***
	(0.01)
LSnudgeXafter	-0.01
	(0.02)
PBnudgeXafter	-0.07***
	(0.02)
after	0.04***
	(0.01)
Constant	0.28***
	(0.02)
N	105,105

*Robust standard errors in parentheses, clustered at individual driver level *** p<0.01, ** p<0.05, *p<0.1, estimated with all controls.*

5.3 Personalized Interventions using Generalized Random Forest Analysis

In a task such as driving that has strong negative externalities (bad drivers causing poor driving conditions for other drivers), it is important to establish that nudges not only make the population better off on average, but also that nudges are not hurting anyone's performance. Further, studying the heterogeneous effects systematically can enable us to personalize the interventions based on important characteristics of individuals. Therefore, we are interested in identifying whether nudges benefit everyone or not or, in other words, whether there is heterogeneity in the treatment effect or not. If we

observe heterogeneity in treatment effect, then a natural implication will be: can we optimize nudges to maximize the treatment effect? Although several methods can be employed to answer these questions, we select generalized random forest or GRF (Athey et al. 2019) for our analysis. We conduct robustness tests with other methods (such as sub-sample analysis) in the robustness section. For brevity of exposition, we provide a short introduction of the GRF method here and refer to Athey et al. (2019) for technical details.

GRF is a generalization of the random forest method and it provides the estimate of *what will be the average treatment effect conditional on different covariates* so it allows one to estimate conditional average treatment effect (CATE). CATE offers a way to personalize interventions, e.g., if nudges have effect only on low-performing users then nudges can be sent to these users only. Methodologically, it preserves several core components of the random forest implementation such as recursive partitioning and random split selection. However, causal forest modifies the averaging of the effect to combine the individual ensemble used in random forest. Instead, it adaptively learns about the weights to be provided to each tree before combining and identifying the average treatment effect. There are several important benefits of using a GRF. First, it uses adaptive weighting, enabling it to accurately identify clusters in a large covariate space. Second, it allows for a large covariate space without compromising on the computational efficiency. The desired consistency and asymptotic properties make it apt for our application to not only identify the heterogeneous treatment effect but to also be able to calculate the 95% confidence intervals. Further, it implements *honesty* to achieve consistent estimation, i.e., separate data is used for two key steps in the estimation process -- growing the trees and estimating the treatment effect. To do so, the algorithm randomly splits the sample data into two equal parts, one half is used for growing the trees and the other half is used for estimation. For ease of reference, we will use similar notation that is used in the original paper describing the GRF method.

Suppose we denote our data as (Y_i, X_i, W_i) where Y_i is the outcome variable, X_i are features or different covariates, W_i denotes treatment assignment and therefore, is a binary variable. We are interested in calculating the average treatment effect τ conditional on $X_i = x$. That is,

$$\tau(x) = \mathbb{E}[\mathrm{Y_i}(1) - \mathrm{Y_i}(0) | \mathrm{X_i} = \mathrm{x}]$$

where, $Y_i(1), Y_i(0)$ denote outcomes observed for treatment and control conditions, respectively. To identify $\tau(x)$ it is essential that treatment assignment satisfies conditional orthogonality to the outcome, i.e., unconfoundedness assumption meaning that treatment is randomly assigned conditional on *x*.

Let as assume that we grow B trees. We index these trees using $b = \{1,2, \ldots., B\}$. We denote the set of training examples falling in the same leaf as x by $L_b(x)$. If $\alpha_{bi}(x)$ is the frequency of i^{th} training example falling into the same leaf as x, then we can represent $\alpha_{bi}(x)$ as

$$\alpha_{bi}(x) = \frac{1(\{X_i \in L_b(x)\})}{|L_b(x)|}.$$

Furthermore, the weights α_i are calculated using the following equation:

$$\alpha_i = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x).$$

Using these weights, the following moment condition is defined to identify the average treatment effect:

$$\sum_{i=1}^{n} \alpha_i \psi_{\tau(x),c(x)}(Y_i, W_i) = 0,$$

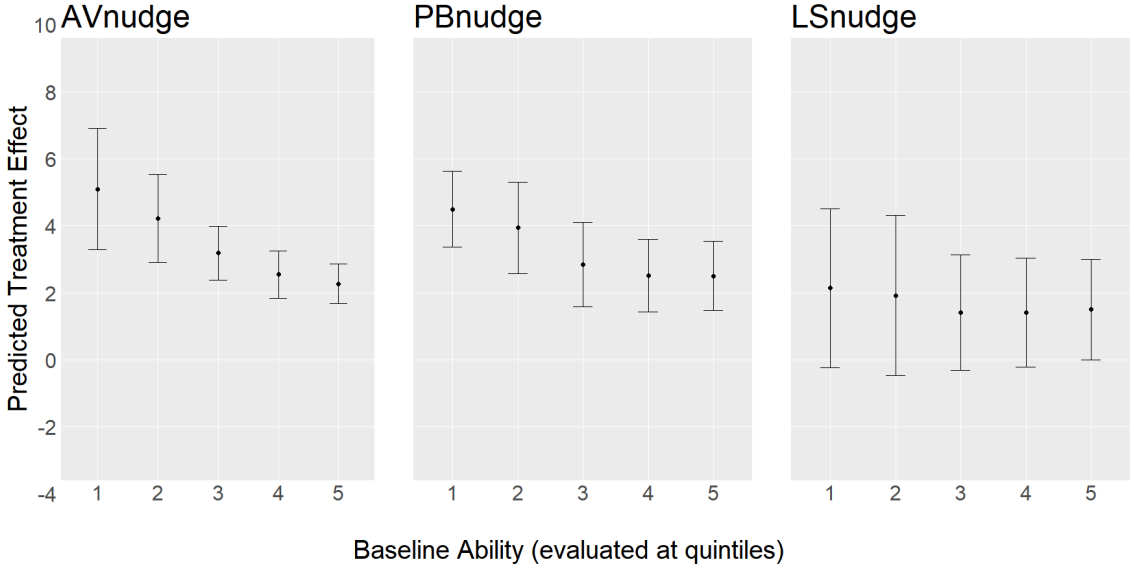
where $\psi_{\tau(x),c(x)}(Y_i, W_i) = (Y_i - c(x) - \tau(x)W_i)\left(1 W_i^T\right)^T$ and $c(x)$ is a nuisance parameter. Finally, solving the above moment condition, the conditional treatment effects are identified as:

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(W_i - \bar{W}_\alpha)\,(Y_i - \bar{Y}_\alpha)}{\left(\sum_{i=1}^{n} \alpha_i(x)(W_i - \bar{W}_\alpha)\,(W_i - \bar{W}_\alpha)^T\right)},$$

where $\bar{W}_\alpha = \sum_{i=1}^{n} \alpha_i W_i$ and $\bar{Y}_\alpha = \sum_{i=1}^{n} \alpha_i Y_i$. It has been shown that $\hat{\tau}(x)$ is asymptotically normal and is consistent with the true treatment effect $\tau(x)$ (Athey and Wager 2019).

To estimate CATE, we use only post-experiment data for our analysis with GRF as the pre-treatment data is fairly balanced meaning there is no difference in performance prior to the launch of experiment so that any treatment effect we observe should be from the after treatment date (we also run robustness tests with pre- and post-data and do not find significant difference in the results). We use **grf** package in R for analysis and estimation of the treatment effect. Although the default number of trees (2,000) were sufficient for the estimation of the average treatment effect, in order to estimate the standard errors accurately we increased the number of trees to 8,000. Increasing the number of trees beyond this number did not change our standard errors significantly.

Figure 3: CATE: Average treatment effect conditioned on baseline ability



For each quintile of baseline ability, the plot represents the mean treatment effect with the respective 95% CI.

We start our analysis by estimating the treatment effect for all three types of nudges. First, we plot the 95% CI of the predicted treatment effects and find that all intervals are strictly above zero for

AVnudge and *PBnudge.* Therefore, our results indicate that these two nudges do not make any driver significantly worse off. However, we observe that for *LSnudge* 95% confidence interval contains both positive and negative values. Therefore, *AVnudge* and *PBnudge* do not make anyone worse off while that is not true for *LSnudge*.

For personalization, we are interested in two parameters – baseline average performance (baseline ability) and average feedback seeking frequency, across which GRF outcomes show large variability in treatment effect[2]. These two characteristics are important for firms in practice. First, firms are interested in targeting individuals with low performance without hurting performance of the well-performing drivers. Second, how frequently users engage with the app through feedback seeking is an important parameter for telematics firms as the main purpose of the app is to interact with users. Therefore, this feedback-seeking behavior is one of the key dimensions for firms in implementing nudge-like interventions that are delivered through the app, i.e., frequent feedback seekers are more likely to take-up the treatment. We operationalized the two variables using individual's pre-treatment latest *todatescore* and percentage of trips after which users have reviewed their detailed feedback by opening the dashboard. This is particularly of importance as high feedback seeking propensity has a negative association with performance on average, as evident from Table 4.

To analyze the heterogeneity in treatment effect due to nudges, we will begin by illustrating the method using a single dimension (i.e., the base line ability) followed by adding the second variable (i.e., feedback seeking). To do so, we first fix all covariates to their median values and estimate the effect along the baseline performance of the drivers (measured by the latest pre-treatment *to-date rating*). First, in Figure 3, we plot the CATE and corresponding 95% confidence intervals conditional on users' ability quintiles. We define ability by the latest pre-treatment *todatescore*. We find that the *AVnudge* and *PBnudge* both have similar effects on performance across drivers (the two left panels). Next, we test whether there is a statistically significant heterogeneous effect using the omnibus method for the test of heterogeneity. Our estimation suggests that there is a significant heterogeneity in treatment effect in case of *AVnudge* as well as *PBnudge* for users with different abilities prior to the launch of the treatment. However, we do not find any heterogeneity in treatment effect in case of *LSnudge.*

Next, we investigate the CATE conditional on two parameters taken together, i.e., interaction of the two variables of interest (ability and feedback seeking). GRF enables us to estimate such interactions and it provides a holistic picture of the CATE variation for the two covariates of interest. We report these results in Figure 4. The scale of the color bar reports the predicted treatment effects, which is the same metric that we use in Figures 3. Evidently, both *AVnudge* and *PBnudge* have similar effects but of varying magnitudes. The first and second panels in Figure 4 show that low performing high feedback

[2]We also considered personalization along a third dimension: variance in the pre-treatment performance score. We find, however, that the two dimensions we selected (feedback seeking and baseline performance) yield the highest benefit in personalization.

seekers benefit the most from these two nudges. Similarly, high performing users who are low feedback seekers benefit the least from these nudges. This observation can be used to personalize nudges which we illustrate by using an example of two categories of drivers (high and low performers) as well as two categories of feedback seekers (high and low feedback seekers). Using the outcome of GRF we find that (results reported in Table 7) the CATE for *PBnudge* is statistically higher than for *AVnudge* for high performance low feedback seeking users (2.87 and 1.21 respectively) while the average effect of CATE for *AVnudge* is higher than the effect of *PBnudge* for low performance but high feedback seekers (5.70 and 4.15 respectively). For other drivers, we do not find any differences between the two nudges. Using these results, we estimate that firms can improve driver performance by nearly 11% through personalization, i.e., sending the low performing high feedback seeking drivers personal average nudge while sending high performing drivers with low propensity to seek feedback personal best nudge.

In Table 7, we further summarize these findings in a simple 2×2 framework that can be used in practice to personalize intervention based on two important dimensions. Nudges are effective for individuals who have room to improve (low performers) and for individuals who are willing to learn (high feedback seekers). Among learners (high feedback seekers), low performers are likely to be more motivated by easy to attain reference points (or goals) and hence, work harder; less challenging goals make personal average nudge more motivating than the personal best nudge. This result shows that firms can benefit from personalizing nudges based on the latest feedback seeking as well as performance of drivers, which is a novel result. In addition to that, our results show that *AVnudge* can be as powerful as *PBnudge* in improving performance, which differs from our anticipation. Finally, we find that there is no benefit of using *LSnudge* over *PBnudge or AVnudge* across any type of drivers.

Table 7: Personalization of nudges

	Low performance		**High performance**	
High feedback frequency	AVnudge >** PBnudge		*AVnudge = PBnudge*	
	5.70	4.15	2.95	3.10
	(0.53)	(0.58)	(0.76)	(1.22)
Low feedback frequency	AVnudge = PBnudge		AVnudge <** PBnudge	
	4.20	4.40	1.21	2.87
	(1.65)	(0.93)	(0.53)	(0.60)

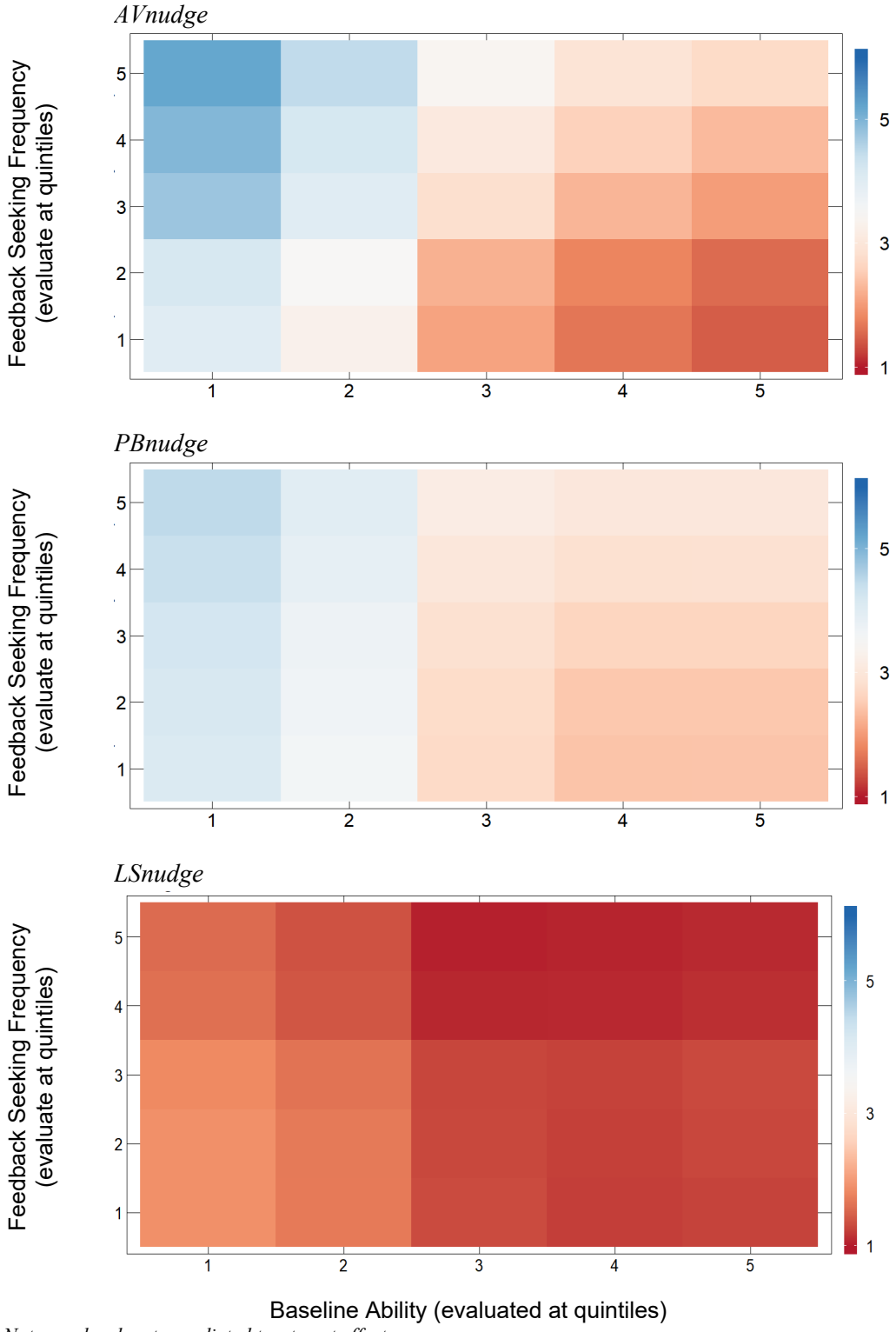
*Numbers denote CATE, standard errors in parentheses clustered for individual drivers, ***p<0.01, ** p<0.05, *p<0.1*

6. The mechanism behind the results

We now turn to the theoretical development to identify some potential mechanisms that are consistent with our data. As argued previously, one possible explanation of our results is goal setting behavior exhibited by drivers using the nudges. Through nudges, we remind drivers about their personal best, personal average, or last performance and how it compares with their current performance. Information on performance promotes goal setting (Locke and Latham 2002), and once users set goals, they tend to

put more effort towards achieving them. As a result, users who set a goal, perform better than those who do not. Further, not all goals are important in improving performance. Challenging but achievable goals are known to improve performance (Bandura 2010). Goals also motivate drivers to make the reference points a desired self-state that drivers try to attain (Austin and Vancouver 1996).

Figure 4: CATE: Interaction of baseline ability and feedback seeking frequency



Note: scales denote predicted treatment effect

All the nudges that we use provide achievable goals as they are based on drivers' own past performance, meaning they have already been achieved before. However, *AVnudge* and *PBnudge* provide users with low variance (var) information to set goals as opposed to *LSnudge:* mathematically, if x is a random variable, var(x) $\geq$ var(mean(x)) and var(x)$\geq$ var(max(x)). Information sent through nudges acts as a reference point that helps in setting goals as well as enables evaluation of performance compared to the reference point. Unfortunately, we do not observe exact goal setting behavior of individuals, however, to test this mechanism, we can use the variation in the reference points which correlates positively with goal setting, meaning if there is low variation in the reference points, users will be able to set goals easily whereas if there is a high variation in the reference points, it will not help users to evaluate their performance or set consistent performance goals.

Table 8: Effect of types of nudges on score for drivers with high and low variation in reference points.

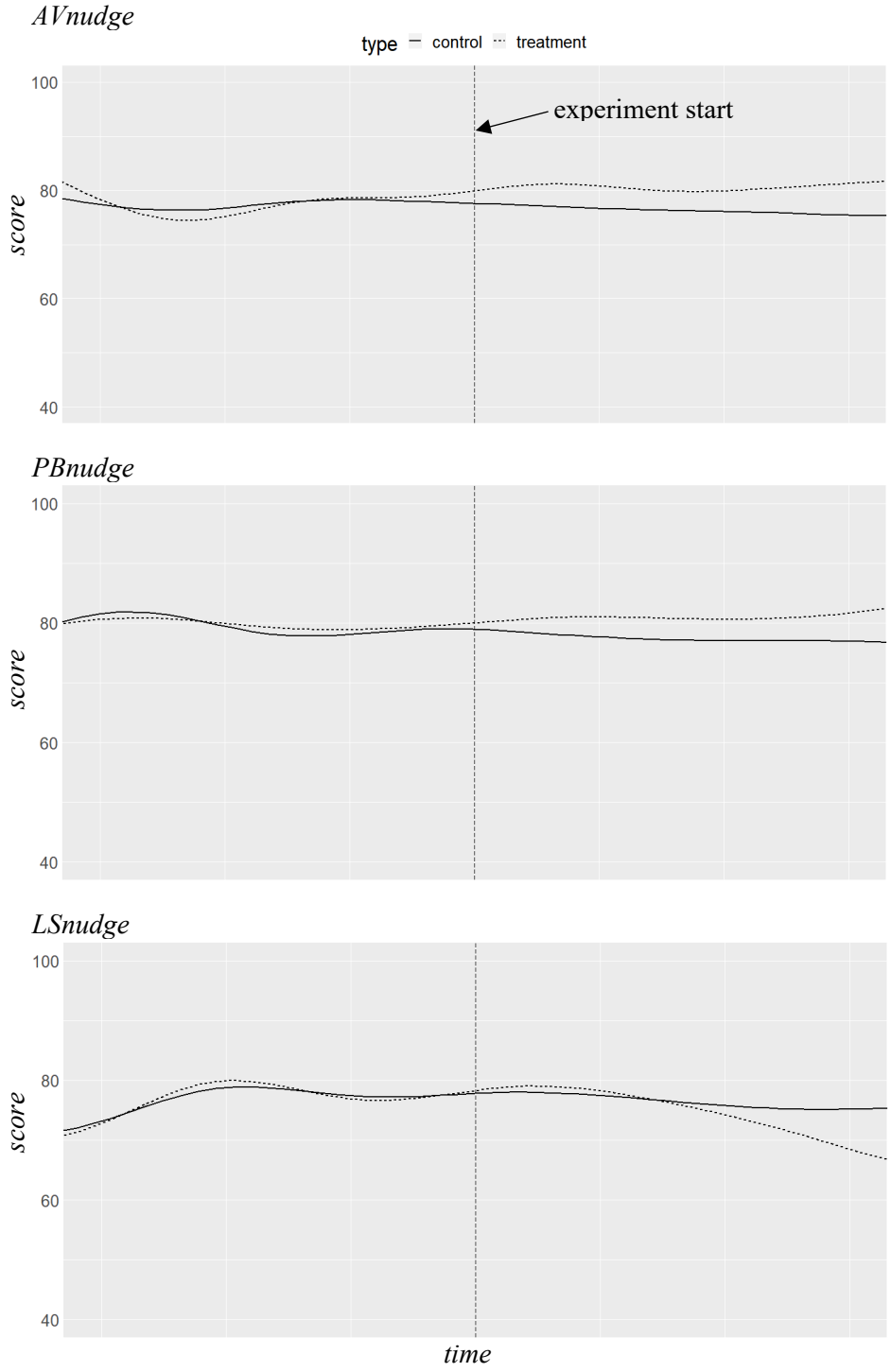
	Dependent variable: score	
	(1)	(2)
AVnudge	4.24	0.17
	(2.81)	(1.69)
Lsnudge	1.50	4.50***
	(2.95)	(1.63)
AvnudgeXafter	2.20	4.84***
	(2.60)	(1.35)
LsnudgeXafter	-3.55	2.46*
	(2.50)	(1.42)
after	-0.58	-1.93
	(2.06)	(1.19)
Constant	83.90***	89.85***
	(2.48)	(1.91)
N	30,372	48,681
SD of reference point	*high*	*low*

*Robust standard errors in parentheses clustered at drivers, *** p<0.01, ** p<0.05, p<0.1, estimated with all controls*

To test this potential mechanism, we calculate the standard deviation of the reference points (personal best, personal average, or latest performance) shared with the drivers using nudges. We then split drivers within each treatment into two groups based on whether the standard deviation of their reference points is above or below the median standard deviation of reference points within treatment. We find that all observations within the *PBnudge* treatment have very small variation as the personal best performance does not vary much with every trip; hence, the split between high and low variation is not meaningful and we do not use it to study the consistency with our proposed mechanism. However, we can use *AVnudge* and *LSnudge,* which have ample variations in the reference points after each trip. Using these sub-samples, we identify the effect of nudges on trip scores. We report the results in Table 8 (columns 1 and 2). We find that there is no effect of *AVnudge* or *LSnudge* on performance when there is a high

variation in the reference points. However, when the variation in reference points is low, we find that there is a positive effect of *both* types of nudges. Thus, our data appears to be consistent with this mechanism.

Figure 5: Synthetic control method plots



7. Robustness Tests

In this section, we conduct several robustness tests to ensure that our estimates are consistent.

7.1 Synthetic Control Method

One of the key assumptions in the DID estimation framework is parallel trends assumption, that is, prior to an intervention, control and treatment groups should exhibit parallel trends. Any violation of this assumption leads to inconsistent estimation of the treatment effect. Although visual inspection can be one way to verify this assumption, there are statistical methods, namely, the synthetic control method, which can address this issue. The method creates artificial control units (or synthetic control units) comparable to the treatment units in the pre-treatment time period. As a result, any differences in post-experiment performance can be attributed to the intervention. Therefore, we can estimate the effect consistently making fundamental assumptions such as random assignment of subjects. In our first robustness test we estimate the effect of different nudges using synthetic control method (Abadie et al. 2010, Xu 2017). The results are plotted in Figure 5. Our results suggest that after creating a synthetic control for the pre-treatment time period, the outcomes are consistent as evident from the plots (even the effect sizes are similar to what we obtain previously ~4.6 points). *PBnudge* and *AVnudge* plots are above the control group post-treatment, while the *LSnudge* seems to follow the control group and eventually drops below the control group in performance. Therefore, our results are not driven by the violation of parallel trends assumption.

Table 9: Effect of nudges on high performing vs low performing drivers

	Dependent variable: score	
	(1)	(2)
AVnudge	-1.47	3.10*
	(1.07)	(1.74)
LSnudge	-1.38	1.51
	(1.13)	(2.33)
PBnudge	-0.70	2.89
	(1.13)	(1.83)
AVnudge×after	3.05**	5.40***
	(1.31)	(1.49)
LSnudge×after	2.21	-1.19
	(1.48)	(1.76)
PBnudge×after	4.01***	4.92***
	(1.34)	(1.60)
after	-3.17***	-1.16
	(1.04)	(1.18)
N	55,300	49,801
Sub-sample	high performers	low performers

*Robust standard errors in parentheses, *** p<0.01, ** p<0.05, p<0.1, estimated with weekend dummies, estimated with all controls.*

7.2 Fixed Effects Model

In our analysis, we have used group level dummies to estimate the effect of nudges on performance.

We also test robustness of our results while accounting for the individual level heterogeneity. To do so, we use individual fixed effect in estimating our regression model. We report these results in Table 9. We find that, accounting for individual-level fixed effects, our results are robust, however, we note smaller effect size.

7.3 Sub-sample Analysis

To test the robustness of our results with GRF, we conduct a robustness test with sub-sample analysis to study the effect of nudges on the score of high vs low performing drivers. To do so, we first define high performance if the latest pre-experiment to-date score is higher than the median value of the *todatescore* of all drivers. Using this classification we run sub sample analysis for the effect of nudges conditional on drivers' performance. We report the results in Table 10. Our analysis are robust and similar to these obtained using GRF. Low-performing drivers have higher effect of personal best and personal average nudges. For instance, the average performance nudge improves the performance of high score drivers by 3.05 points whereas it improves performance for low-performing drivers by 5.40 points, on an average.

7.4 Additional analysis – drivers who join the program after treatment

As mentioned in the experiment design section, we allowed drivers to join the program even after the launch of treatment. We did this to ensure that our results are robust. Our industry partners supported us by not limiting the number of drivers enrolled. The 540 drivers as described in Table 11 were randomly assigned to the three treatment and control group but with a different probability of random assignment this time – 40% for control, 20% for treatment[3]. All treatment drivers started receiving same nudge messages as the previous 529 drivers. Naturally, we could not collect baseline data for these drivers[4]. Analysis of these drivers reveals that our results are similar to analysis of the main 529 drivers, that is, personal best and personal average nudges have a positive effect on trip performance while last score nudge does not. Therefore, our results remain qualitatively unchanged.

Table 10: Number of drivers and trips across groups joining after the experiment started

Drivers		*Control*	*Treatment*			*Total*
			Personal best nudge (PBnudge)	*Last score nudge (LSnudge)*	*Personal average nudge (AVnudge)*	
Onlyafter	*#drivers*	223	109	106	102	540
	#trips	6,130	5,994	6,218	5,939	24,281

8. Discussion and Conclusion

In this paper, we provide evidence of the effectiveness of a novel intervention to improve driving

[3] The industry partner did not agree for an equal probability of assignment due to the concern of attrition.

[4] Due to the limitation in the platform, nudges were sent as soon as a driver is assigned to a treatment group. Therefore, we could not stop sending the nudges to these drivers to collect baseline data.

performance. Unlike financial incentives, this intervention is inexpensive. Nudges have been used in many other contexts including road safety (e.g., painting stripes on an S curve to nudge drivers, Thaler and Sunstein (2008), p38), but, to the best of our knowledge, we are the first to conduct a field experiment and show the effectiveness of nudges in improving driving behavior using telematics technology. Driving is core to many supply chain processes and any small changes in driving can have substantial effect in supply chain efficiencies and costs. Nudges have been studied for decades, however, we use them in a context where the task at hand is quite complex (driving) and the behavior is difficult to change.

Our results provide practitioners with a novel and inexpensive way to improve performance through nudges. Using novel methodology in our analysis we provide evidence of the benefits achieved through personalizing these nudges. We show that, instead of using one type of nudge for all users, firms can benefit from personalizing the nudges based on the current performance and feedback-seeking propensity of the drivers. In Table 7, we also provide a framework for implementation of a nudge-like personalized interventions for improving performance. We provide evidence of a mechanism motivated from goal-setting theory and find it to be consistent with our data.

One point worth highlighting is that many firms provide users with the last trip performance while our results indicate that providing such feedback (which has high variance) may not have the desired effect on driving performance, the finding that is in line with a related study in the same context (Choudhary et al. 2018). Finally, our research informs the nudge literature to provide evidence that testing of nudges in a field setting is an important part of policy decision (Sunstein 2014) as not all nudges work equally in improving performance.

Although we focus on driving in this paper, our results have wider applications in several areas beyond driving. For example, in industries such as education and logistics, feedback is regularly provided for performance management and therefore personalized nudges can be effective in these settings.. For instance, our results imply that we can improve performance in education using personalized nudges based on a student's performance and eagerness to learn (for example, measured by student's interaction with the online feedback/grading platform). In logistics, many warehousing firms now employ IoT devices to provide information on picking orders and real-time performance. Nudges can be implemented to motivate different types of employees based on their performance and how often they review their performance.

There are many follow-up questions that this study provides a path to, for instance, what is the optimal nudge schedule, what is an optimal nudge message, etc.? These natural questions are important for both academics as well as practitioners and merit separate studies. Our results on the personalization of nudges lay out a path for future inquiry on the dynamic effect of nudges, specifically, how drivers will respond to the nudges when they are dynamically adjusted based on their most recent trips (e.g., past week performance and feedback seeking behavior). It will be also interesting to study the effect of nudges with different frequency as well as different types of feedback such as social comparison. Lastly,

our setup does not allow us to study the long-term effect of nudges on performance. We believe our results will motivate further research in this area.

References

Abadie A, Diamond A, Hainmueller J (2011) Synth : An R Package for Synthetic Control Methods in Comparative Case Studies. *J. Stat. Softw.* 42(13):1–17.

Abadie A, Diamond A, Hainmuellera J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Am. Stat. Assoc.* 105(490):493–505.

Allcott H (2011) Social norms and energy conservation. *J. Public Econ.* 95(9–10):1082–1095.

Anderson A, Green EA (2018) Personal bests as reference points. *Proc. Natl. Acad. Sci.* 115(8):1772–1776.

Anderson A, Robinson DT (2017) Self-Awareness, Financial Advice and Retirement Savings Decisions. *SSRN Electron. J.*

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *Ann. Stat.* 47(2):1179–1203.

Athey S, Wager S (2019) Estimating Treatment Effects with Causal Forests: An Application. :1–15.

Austin JT, Vancouver JB (1996) Goal constructs in psychology: Structure, process, and content. *Psychol. Bull.* 120(3):338.

Balafoutas L, Fornwagner H, Sutter M (2018) Closing the gender gap in competitiveness through priming. *Nat. Commun.* 9(1):4359.

Bandura A (2010) Self-efficacy. *Corsini Encycl. Psychol.*:1–3.

BBC (2014) When conventional road signs have no effect, designers are turning to increasingly clever ways to subconsciously make drivers slow down or pay attention. http://www.bbc.com/future/story/20140417-road-designs-that-trick-our-minds.

Bendoly E (2013) Real-time feedback and booking behavior in the hospitality industry: Moderating the balance between imperfect judgment and imperfect prescription. *J. Oper. Manag.* 31(1–2):62–71.

Bhattacharya J, Garber AM, Goldhaber-Fiebert J (2017) Nudges in Exercise Commitment Contracts: A Randomized Trial. *SSRN Electron. J.*

Bolderdijk JW, Knockaert J, Steg EM, Verhoef ET (2011) Effects of Pay-As-You-Drive vehicle insurance on young drivers' speed choice: Results of a Dutch field experiment. *Accid. Anal. Prev.* 43(3):1181–1186.

Bolton GE, Katok E (2008) Learning by Doing in the Newsvendor Problem: A Laboratory Investigation of the Role of Experience and Feedback. *Manuf. Serv. Oper. Manag.* 10(3):519–538.

Burns EC, Martin AJ, Collie RJ (2019) Understanding the role of personal best (PB) goal setting in students' declining engagement: A latent growth model. *J. Educ. Psychol.* 111(4):557–572.

Camilleri AR, Larrick RP (2014) Metric and scale design as choice architecture tools. *J. Public Policy Mark.* 33(1):108–125.

Charlier C, Guerassimoff G (2018) Under Pressure! Nudging Electricity Consumption within Firms: Feedback from a Field Experiment. *GREDEG Work. Pap. 2019-18*.

Chen Y, Lu F, Zhang J (2017) Social Comparisons, Status and Driving Behavior. *J. Public Econ.* 155:11–20.

Choudhary V, Shunko M, Netessine S (2018) Does Real-Time Feedback Make You Try Less Hard?: A Study of Automotive Telematics. *SSRN Electron. J.*

Cramer J, Krueger AB (2016) Disruptive Change in the Taxi Business: The Case of Uber. *Am. Econ. Rev. Pap. Proc.* 106(5):177–182.

Cronqvist H, Thaler RH (2004) Design Choices in Privatized Social-Security Systems: Learning from the Swedish Experience. *Am. Econ. Rev.* 94(2):424–428.

Croson R, Donohue K (2006) Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information. *Manage. Sci.* 52(3):323–336.

DeJoy DM (1992) An examination of gender differences in traffic accident risk perception. *Accid. Anal. Prev.* 24(3):237–246.

Donohue K, Özer Ö, Zheng Y (2019) Behavioral Operations: Past, Present, and Future. *Manuf. Serv. Oper. Manag.* (November):1–12.

Du N, Shelton S, Whittington R (2012) Does Supplementing Outcome Feedback with Performance Feedback Improve Probability Judgments? *Int. J. Financ. Res.* 3(4):19–32.

FleetAnswers (2018) Strategies to sustain good driving behavior in the long-run. *FleetAnswers*.

Fredrickson BL, Kahneman D (1993) Duration Neglect in Retrospective Evaluations of Affective Episodes. *J. Pers. Soc. Psychol.* 65(1):45–55.

Gaker D, Zheng Y, Walker J (2010) Experimental Economics in Transportation: A Focus on Social Influences and the Provision of Information. *Univ. Calif. Transp. Cent.* UCTC-FR-20.

Grossman Z, Owens D (2012) An unlucky feeling: Overconfidence and noisy feedback. *J. Econ. Behav. Organ.* 84(2):510–524.

Habyarimana J, Jack W (2015) Results of a large-scale randomized behavior change intervention on road safety in Kenya. *Proc. Natl. Acad. Sci. U. S. A.* 112(34):E4661–E4670.

Helweg-Larsen M, Shepperd JA (2001) Do Moderators of the Optimistic Bias Affect Personal or Target Risk Estimates? A Review of the Literature. *Personal. Soc. Psychol. Rev.* 5(1):74–95.

Ibanez MR, Staats BR (2018) Behavioral Empirics and Field Experiments. *Handb. Behav. Oper.* (John Wiley & Sons, Ltd), 121–147.

Kagan E, Leider S, Lovejoy WS (2018) Ideation – Execution Transition in Product Development : An Experimental Analysis. *Manage. Sci.* 64(5):2238–2262.

Kim S hee, Tong J, Peden C (2018) The Effects of Occupancy Information Hurdles and Physician Admission Decision Noise on Hospital Unit Utilization. *SSRN Electron. J.* (2012).

Knechtle B, Wirth A, Knechtle P, Zimmermann K, Kohler G (2009) Personal best marathon performance is associated with performance in a 24-h run and not anthropometry or training

volume. *Br. J. Sports Med.* 43(11):836–839.

Kremer M, Moritz B, Siemsen E (2011) Demand Forecasting Behavior: System Neglect and Change Detection. *Manage. Sci.* 57(10):1827–1843.

Locke EA, Latham GP (2002) Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *Am. Psychol.* 57(9):705–717.

Lurie NH, Swaminathan JM (2009) Is Timely Information Always Better? The Effect of Feedback Frequency on Decision Making. *Organ. Behav. Hum. Decis. Process.* 108(2):315–329.

Markus MM, Niederle M, Rosenblat TS (2014) *Managing Self-Confidence*

Marteau TM, Ogilvie D, Roland M, Suhrcke M, Kelly MP (2011) Judging Nudging: Can Nudging Improve Population Health? *BMJ* 342(Jan):d228–d228.

Martin AJ (2006) Personal bests (PBs): A proposed multidimensional model and empirical analysis. *Br. J. Educ. Psychol.* 76(4):803–825.

Massey C, Wu G (2005) Detecting Regime Shifts: The Causes of Under- and Overreaction. *Manage. Sci.* 51(6):932–947.

PennMedNews (2018) Penn and CHOP Receive $1.84 Million to Study Ways to Curb Cell Phone Use while Driving. https://www.pennmedicine.org/news/news-releases/2018/november/penn-and-chop-team-receives-grant-to-study-best-practices-for-curbing-cell-phone-use-while-driving.

Roy MM, Liersch MJ (2013) I Am a Better Driver Than You Think: Examining Self- Enhancement for Driving Ability. *J. Appl. Soc. Psychol.* 43(8):1648–1659.

Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V (2007) The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychol. Sci.* 18(5):429–434.

Soleymanian M, Weinberg CB, Zhu T (2019) Sensor Data and Behavioral Tracking: Does Usage-Based Auto Insurance Benefit Drivers? *Mark. Sci.* 38(1):21–43.

Song H, Tucker AL, Murrell KL, Vinson DR (2018) Closing the Productivity Gap: Improving Worker Productivity Through Public Relative Performance Feedback and Validation of Best Practices. *Manage. Sci.* 64(6):2628–2649.

Stillwater T, Kurani KS (2013) Drivers discuss ecodriving feedback: Goal setting, framing, and anchoring motivate new behaviors. *Transp. Res.* 19(Part F: Traffic Psychology and Behavior):85–96.

Sunstein CR (2014) Nudging: A Very Short Guide. *J. Consum. Policy* 37(4):583–588.

Sunstein CR (2017) Nudges that fail. *Behav. Public Policy* 1(1):4–25.

Szaszi B, Palinkas A, Palfi B, Szollosi A, Aczel B (2018) A Systematic Scoping Review of the Choice Architecture Movement: Toward Understanding When and Why Nudges Work. *J. Behav. Decis. Mak.* 31(3):355–366.

Tayur S (2017) OM forum - An essay on operations management. *Manuf. Serv. Oper. Manag.* 19(4):526–533.

Tayur S, Kush J, Aven B (2019) Nudge(able): Field and Experimental Studies in Organ Donation. *Work.*

Pap.

Terwiesch C, Olivares M, Staats BR, Gaur V (2019) A Review of Empirical Operations Management over the Last Two Decades. *Manuf. Serv. Oper. Manag.* (November):1–13.

Thaler RH, Benartzi S (2004) Save More Tomorrow ™: Using Behavioral Economics to Increase Employee Saving. *J. Polit. Econ.* 112(1):164–187.

Thaler RH, Sunstein CR (2008) *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Yale University Press).

WHO (2018) *Global Status Report on Road Safety*

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* Second. (MIT Press).

Xu Y (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Polit. Anal.* 25(1):57–76.

Yang N, Long Lim Y (2018) Temporary Incentives Change Daily Routines: Evidence from a Field Experiment on Singapore's Subways. *Manage. Sci.* 64(7):3365–3379.

Yao WX, Fischman MG, Wang YT (1994) Motor Skill Acquisition and Retention as a Function of Average Feedback, Summary Feedback, and Performance Variability. *J. Mot. Behav.* 26(3):273–282.