Transportation Research Part F 71 (2020) 198–219

Contents lists available at ScienceDirect

Transportation Research Part F

journal homepage: www.elsevier.com/locate/trf

App-based feedback on safety to novice drivers: Learning and monetary incentives

Stefanie Peer [a,*], Alexander Muermann [a,b], Katharina Sallinger [c]

[a] Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria
[b] Dolphin Technologies GmbH, Stella-Klein-Löw-Weg 11, 1020 Vienna, Austria
[c] Vienna Graduate School of Finance (VGSF), Welthandelsplatz 1, 1020 Vienna, Austria

ARTICLE INFO

Article history:
Received 10 January 2020
Received in revised form 30 March 2020
Accepted 7 April 2020

Keywords:
Traffic safety
Monetary incentives
App-based feedback
Novice drivers
Field experiment
Telematics

ABSTRACT

An over-proportionally large number of car crashes is caused by novice drivers. In a field experiment, we investigated whether and how car drivers who had recently obtained their driving license reacted to app-based feedback on their safety-relevant driving behavior (speeding, phone usage, cornering, acceleration and braking). Participants went through a pre-measurement phase during which they did not receive app-based feedback but driving behavior was recorded, a treatment phase during which they received app-based feedback, and a post-measurement phase during which they did not receive app-based feedback but driving behavior was recorded. Before the start of the treatment phase, we randomly assigned participants to two possible treatment groups. In addition to receiving app-based feedback, the participants of one group received monetary incentives to improve their safety-relevant driving behavior, while the participants of the other group did not. At the beginning and at the end of experiment, each participant had to fill out a questionnaire to elicit socio-economic and attitudinal information.

We conducted regression analyses to identify socio-economic, attitudinal, and driving-behavior-related variables that explain safety-relevant driving behavior during the pre-measurement phase and the self-chosen intensity of app usage during the treatment phase. For the main objective of our study, we applied regression analyses to identify those variables that explain the potential effect of providing app-based feedback during the treatment phase on safety-relevant driving behavior. Last, we applied statistical tests of differences to identify self-selection and attrition biases in our field experiment.

For a sample of 130 novice Austrian drivers, we found moderate improvements in safety-relevant driving skills due to app-based feedback. The improvements were more pronounced under the treatment with monetary incentives, and for participants choosing higher feedback intensities. Moreover, drivers who drove relatively safer before receiving app-based feedback used the app more intensely and, ceteris paribus, higher app use intensity led to improvements in safety-related driving skills. Last, we provide empirical evidence for both self-selection and attrition biases.

* Corresponding author.
E-mail address: stefanie.peer@wu.ac.at (S. Peer).

https://doi.org/10.1016/j.trf.2020.04.005
1369-8478/

1. Introduction

Young drivers (typically defined as drivers younger than 24 years old) have a higher risk of being involved in car crashes, with the accident risk being two to three times higher than that of older drivers (SafetyNet, 2009; Statistik Austria, 2017). In Europe, about a fifth of all deaths in the age group of 14 to 24 years can be attributed to road accident fatalities (Laiou et al., 2010). Young drivers are usually novice drivers, as the minimum age for obtaining the driving license is commonly 18 years, and in some countries less. Hence, young drivers only had limited time to gain driving experience. Conversely, novice drivers tend to be young, as most persons obtain their driving license close to when they exceed the minimum age requirement, rather than later in life.

A major cause for car crashes among novice drivers is the lack of driving experience, which leads, among others, to recognition errors and inadequate speeds (Braitman, Kirley, McCartt, & Chaudhary, 2008; Curry, Hafetz, Kallan, Winston, & Durbin, 2011; McKnight & McKnight, 2003; Neyens & Boyle, 2007). Curry et al. (2011), for example, used US data on car crashes among teen drivers and found that about half of the crashes with severe consequences happened because the drivers did not engage in adequate surveillance, got distracted, or were driving too fast, whereas the driving environment, the vehicle, and aggressive driving played a relatively minor role. Similarly, Klauer et al. (2014) found that, unlike with experienced drivers, the accident risk of newly licensed drivers below the age of 17 years increased with the performance of secondary tasks, including the use of cell phones.

Besides a lack of driving experience contributing to accident risk, lifestyle choices, susceptibility to social influence, and immaturity associated with young age tend to induce more risky driving, in particular among male drivers (e.g. Harré, Field, & Kirkwood, 1996; Simons-Morton et al., 2011; Sarma, Carey, Kervick, & Bimpeh, 2013).[1] Nighttime driving, peer passengers, and other in-car distractions (in particular from cell phone use) have been identified as specific risk factors for young, novice drivers (e.g. Williams, 2003; Simons-Morton et al., 2011; Curry et al., 2011; Sarma et al., 2013). The higher accident risk among young, novice drivers has also been partially attributed to neurophysiological factors, in particular the fact that the pre-frontal cortex, which controls executive functions, matures only around the age of 25 years (Dahl, 2008; Sarma et al., 2013). This may contribute to an increased propensity for risk-taking, sensation seeking and trait impulsiveness, and a lower ability to estimate and manage risks, self-regulate emotions, and focus attention among young drivers.

Traditionally, driving lessons and other forms of supervised driving have been employed in order to improve the safety-relevant driving skills of those in process of obtaining and of those who have recently obtained their driving license. However, they tend to involve substantial monetary costs (if lessons are taken with driving instructors) and/or time costs (if parents or other adults provide supervision). Recently, new technologies have been explored as an instrument for improving safety-relevant driving skills of novice drivers (see Kervick (2016) for an overview): in-vehicle recorders and other devices capable of using GPS (such as smartphones) can be used to collect information related to driving behavior (e.g., speeding, braking, accelerating) and subsequently provide drivers with feedback regarding their behavior. Such data collected from moving objects is often referred to as telematic data. Unlike with driving lessons, scaling up the provision of telematics-based feedback to additional drivers tends to be low-cost (especially if pre-existing devices like smartphones are used).[2]

Most studies providing telematics-based feedback to novice drivers have concluded that safety-relevant driving skills only improve if participation is compulsory, or if participants are offered some type of incentive (Lahrmann et al., 2012; Lotan, 2014; Musicant & Lotan, 2016; Young, Regan, Triggs, Jontof-Hutter, & Newstead, 2010). So far, a direct comparison between the effectiveness of feedback via telematic information and an additional provision of incentives has not been investigated for novice drivers. This is the main aim of our study.

1.1. Using telematics-based feedback to improve safety-relevant driving skills

In recent years, technological advances have increased the quality and real-time (or near-time) availability of telematic information, rendering safety-relevant monitoring and feedback features more detailed, user-friendly, and inexpensive (e.g. Johnson & Trivedi, 2011; Paefgen, Kehr, Zhai, & Michahelles, 2012; Toledo, Musicant, & Lotan, 2008). These qualities are underlined by the willingness of professional drivers to accept and react to technology-based feedback (Roetting, Huang, McDevitt, & Melton, 2003; Huang, Roetting, McDevitt, Melton, & Smith, 2005), and most importantly, by observed reductions in accident numbers in the range of 15 to 30% (Lehmann & Cheale, 1998; Wouters & Bos, 2000).

In the context of traffic safety, telematic information has been used so far to measure a wide number of indicators that reflect (un) safe driving behavior. These include speeding behavior, acceleration, cornering, braking, smartphone use, close following, signal use, and brake position. An insightful overview is provided by Horrey, Lesch, Dainoff, Robertson, and Noy (2012). A number of studies derived compound indicators for aggressive or unsafe driving styles, usually based on (longitudinal and lateral) acceleration and deceleration behavior and resulting gravitational force (g-force) events caused for instance by hard breaking and sharp turns (Farmer, Kirley, & McCartt, 2010; Meseguer, Calafate, Cano, & Manzoni, 2013;

[1] In-depth analyses of other socio-economic and attitudinal characteristics of novice drivers on traffic safety can for instance be found in Holte (2012), Taubman-Ben-Ari and Yehiel (2012), and Ulleberg and Rundmo (2003).

[2] Time costs might still persist if parents are involved (Farah et al., 2014; Guttman & Lotan, 2011; Lerner et al., 2010; Toledo, Lotan, Taubman-Ben-Ari, & Grimberg, 2012).

Musicant & Lotan, 2016; Paefgen et al., 2012; Simons-Morton et al., 2011; Soriguera & Miralles, 2016; Toledo et al., 2008; Vaiana et al., 2014). Outside the domain of traffic safety, telematic information has been used in the context of fostering sustainable driving (Beloufa et al., 2017; Beusen et al., 2009; Tulusan, Staake, & Fleisch, 2012; Wåhlberg &, 2007), and for identification of external road conditions such as potholes and bumps on the road (e.g. Fazeen, Gozick, Dantu, Bhukhiya, & González, 2012; Islam, Buttlar, Aldunate, & Vavrik, 2014; Allouch, Koubâa, Abbes, & Ammar, 2017).

Telematic information has traditionally been derived from in-vehicle data recorders, which are usually hard-wired to a vehicle and make use of accelerometers, cameras, global positioning systems (GPS), radar sensors, lane trackers, and other information from vehicle's on-board diagnostics system (e.g. Lehmann & Cheale, 1998; Toledo et al., 2008; Farmer et al., 2010; Farah et al., 2013). More recently, smartphones have gained importance in collecting telematic information, often via dedicated apps (e.g. Johnson & Trivedi, 2011; Meseguer et al., 2013; Vaiana et al., 2014; Kervick, 2016; Soriguera & Miralles, 2016; Botzer, Musicant, & Perry, 2017). Such information is usually derived from smartphones' internal GPS, their (three-axes) accelerometer sensors, and, more rarely, also their cameras (e.g. Chuang, Bala, Bernal, Paul, & Burry, 2014).

In-vehicle recorders generate data that track a specific vehicle over time, while smartphones generate data that track a specific driver over time. Another difference between the two datasources is that a smartphone app can usually be activated or deactivated by the driver, whereas deactivating an in-vehicle data recorder tends to be less straightforward. In contrast to smartphone apps, some in-vehicle data recorders are even capable of intervening in the actual driving, for instance by imposing a maximum speed, or engaging in emergency braking (Kervick, Hogan, O'Hora, & Sarma, 2015). Smartphone-based systems can only interfere with smartphone use; for instance some apps may block phone use during driving (Vegega, Jones, & Monk, 2013).

Telematic data generated from both in-vehicle recorders and smartphone apps can be used to monitor driving behavior and to provide feedback to drivers. In some instances, monitoring may be the primary goal, and feedback might only occur in exceptional instances (e.g. after a car crash has happened). In others, immediate feedback (e.g. in the case of a potential collision, see Botzer et al. (2017)) may be central. Our focus is on instances where feedback is provided at a regular basis (e.g., after each trip, weekly, monthly) rather than in real-time (e.g. through blinking lights or audio signals), with the aim to improve safety-relevant driving skills (e.g. Toledo et al., 2008; Simons-Morton et al., 2011; Farah et al., 2013). There is evidence that even the mere presence of a monitoring device may lead to improvements in safety-relevant driving patterns (Horrey et al., 2012; Lehmann & Cheale, 1998). For young drivers, feedback may not only be communicated to the drivers themselves, but also to their parents (e.g. Musicant & Lotan, 2016).

Most existing studies do not observe whether and to which extent drivers actually made use of the telematics-based feedback, i.e. to which extent and how drivers interact with the feedback device (e.g. smartphone app). We refer to the extent to which drivers look up telematics-based feedback and the type of feedback that drivers decide to look up as "self-chosen feedback intensity". Literature from other fields such as self-regulated (online) learning environments suggests that the self-chosen feedback intensity is positively related to the motivation level of the participant (see for instance De Barba, Kennedy, & Ainley (2016) who found that both active participation and motivation drive the participants' performance in massive open online courses (MOOCs)). In this study, we are able to observe and control for the effect of self-chosen feedback intensity when explaining improvements in safety-relevant driving skills.

1.2. Traffic safety and incentives

The improvement of safety-relevant driving skills is usually not a sufficient reason for using a driving feedback device, in particular for young drivers (Guttman & Gesser-Edelsburg, 2011; Musicant & Lotan, 2016). Hurst (1980) was among the first to discuss the use of rewards for promoting safe driving, hence trying to overcome the lack of intrinsic motivation by fostering extrinsic motivation. In more recent empirical studies, incentivizing traffic safety either took place – indirectly – by incentivizing the installation (e.g. Chorlton, Hess, Jamson, & Wardman, 2012) or by punishing the de-installation of monitoring and feedback devices (e.g. Etzioni, Erev, Ishaq, Elias, & Shiftan, 2017), or – directly – by incentivizing behaviors linked to safe driving such as adherence to speed limits, no nighttime driving, avoidance of extreme maneuvers, seatbelt usage, or a reduction in overall distance driven within a certain time span (e.g. Greaves & Fifer, 2010; Mazureck & Hattem, 2006; Reagan, Bliss, Van Houten, & Hilton, 2013; Musicant & Lotan, 2016). The technologies that became available in recent years have improved in particular the possibilities for direct incentivization, for instance in the form of pay-as-you-drive insurance schemes, which are increasingly offered by insurance companies (e.g. Bolderdijk, Knockaert, Steg, & Verhoef, 2011; Hultkrantz & Lindberg, 2011; Dijksterhuis et al., 2015; Geyer, Kremslehner, & Muermann, 2019). Most studies involving behavior-dependent ("direct") incentives found significant improvements in safety-relevant driving skills (see Elvik (2014) for an overview).

These improvements may, however, not be generalizable to a wider population, as they are usually based on studies with voluntary participation (see Lahrmann et al. (2012) for an informative discussion on the difficulties to recruit drivers for a traffic safety experiment). Resulting self-selection and attrition biases, leading to an over-representation of motivated participants and hence an overestimation of the improvements, are usually ignored in the analyses (Elvik, 2014). There is also evidence that the effectiveness of the incentives wears off, as the novelty of an incentive scheme wanes (Musicant & Lotan, 2016).

1.3. Feedback vs. incentives

Two studies have so far contrasted the effectiveness of feedback and incentives in the domain of traffic safety. Both Reagan et al. (2013) and Mullen, Maxwell, and Bedard (2015) distinguished (in an experimental setup) between groups of drivers who received only feedback, only (behavior-dependent) incentives, and feedback plus incentives. In both studies feedback was limited to speeding: the study of Mullen et al. (2015) took place in the laboratory (hence, representing only hypothetical behavior), whereas Reagan et al. (2013) used an in-vehicle recorder to provide real-time feedback during the journey (by warning drivers when they exceeded the local speed limit). Both studies found that incentives are a stronger driver for improvements in safety-relevant driving skills than feedback: Reagan et al. (2013) found that when drivers were exposed to both feedback and incentives, reductions in speeding resembled those that resulted when only incentives were provided. Mullen et al. (2015) found that drivers who obtained both feedback and incentives exhibited the least amount of speeding; however, they also showed that when only feedback was provided, the observed speeds were similar to the speeds observed among members of the control group (i.e. individuals who were neither exposed to feedback nor incentives).

1.4. Objectives

In this paper, we aim at testing whether safety-relevant driving skills among novice drivers (who have received their driving license in the past two years) can be improved by providing detailed feedback after each trip (including map-based overviews). More specifically, we investigate whether feedback alone is sufficient to foster learning (due to intrinsic motivation), or whether (behavior-dependent) incentives (and hence extrinsic motivation) are necessary for (safety-relevant) driving skills to improve significantly. We make use of a state-of-the art smartphone app with multiple indicators for safety-relevant driving skills (not only speeding, but also phone usage, cornering, braking and acceleration behavior). We can thus test whether the results obtained by Reagan et al. (2013) and Mullen et al. (2015) hold also for novice drivers and for a more general set of indicators for safety-relevant driving skills (rather than speeding only). Moreover, we aim at shedding light on the moderating role of socio-economic and attitudinal variables, technological operability (iOS vs. Android), biases resulting from participation and dropout decisions, and the feedback intensity chosen by participants.

2. Material and methods

2.1. Participants

In order to participate in the field experiment, three requirements had to be fulfilled by prospective participants:

- having received their driving license in the past two years,
- driving a car regularly,
- owning a suitable smartphone.

Moreover, for our analyses, we only considered drivers for whom we observed at least 10 trips before the start of the experimental treatment. The resulting dataset relevant for the investigation of the impact of app-based feedback on safety-relevant driving skills comprises 130 drivers. The average age of that sample is 20.5 years (sd: 4.6 years) and 45.0% are female. 44.0% live in areas with less than 5000 inhabitants and 54.1% have a monthly net income of more than 2500 Euro (see also Table 1). Overall, over the course of the field experiment, we recorded 12,437 trips (5153 in the pre-measurement, 6746 in the treatment phase, and 538 in the post-measurement phase), covering 220,510.2 km. In A, we provide a flow diagram that filters down from the recruitment to the participants (and various sub-groups).

2.2. Experimental design

The field experiment covered in this paper aims at investigating how, whether, and under which conditions novice drivers react to app-based feedback on driving skills. Their skills were measured using telematic data generated by a dedicated smartphone app (*eMentoring app*), which the participants had to install for the duration of the field experiment (in addition to an on-board diagnostic (OBD) dongle that was used to identify relevant car trips). The participants went through three distinct treatment phases over the duration of the field experiment:

- a **pre-measurement phase** of maximum 6 weeks, in which no (app-based) feedback on (safety-relevant) driving skills was provided,
- a 2-month **treatment phase**, in which app-based feedback on (safety-relevant) driving skills was provided to the participants. Participants were able to see their overall score as well as sub-scores (smartphone usage, speed, acceleration, cornering, braking, speed in dangerous areas), both at the aggregate as well as at the trip level,
- a **post-measurement phase** of maximum one month, in which no (app-based) feedback on (safety-relevant) driving skills was provided.

Table 1
Descriptive statistics for participants and non-participants (test: univariate p-values from χ^2 or Fisher-exact tests comparing participants with at least 10 trips during pre-measurement (n_0) to non-participants (n_2); effect sizes are Cramer's V).

		Participants				Test	
		>= 10	< 10	Non-part.	All	Part. vs. Non-Part.	
Variable	Levels	% (n_0)	% (n_1)	% (n_2)	% (n_{all})	Effect Size	P-value
Age	< 20 years	73.8 (130)	82.6 (109)	79.7 (850)	79.2 (1089)	0.04	0.13
	>= 20 years	26.1 (130)	17.4 (109)	20.4 (850)	20.8 (1089)		
Gender	Male	55.4 (130)	55.0 (109)	49.9 (863)	51.1 (1102)	0.03	0.26
	Female	44.6 (130)	45.0 (109)	50.1 (863)	48.9 (1102)		
Highest education level	Compulsory education	20.0 (130)	16.5 (109)	14.2 (863)	15.2 (1102)	0.09	0.08
	Vocational training	24.6 (130)	9.2 (109)	23.9 (863)	22.5 (1102)		
	High school – no degree	26.1 (130)	26.6 (109)	22.4 (863)	23.2 (1102)		
	High school degree	26.1 (130)	45.9 (109)	37.3 (863)	36.8 (1102)		
	University	3.1 (130)	1.8 (109)	2.2 (863)	2.3 (1102)		
Monthly net income	<= 2500 Euro	53.1 (130)	45.9 (109)	60.2 (863)	58.0 (1102)	0.05	0.13
	> 2500 Euro	46.9 (130)	54.1 (109)	39.8 (863)	42.0 (1102)		
Population of home municipality	< 5000	46.9 (130)	44.0 (109)	49.6 (863)	48.7 (1102)	0.02	0.57
	>= 5000	53.1 (130)	56.0 (109)	50.4 (863)	51.3 (1102)		
Smartphone platform	Android	49.6 (129)	60.7 (84)		54.0 (213)		
	iOS	50.4 (129)	39.3 (84)		46.0 (213)		
Driving education completed	0–12 months ago	56.9 (130)	55.0 (109)	55.2 (863)	55.4 (1102)	0.009	0.78
	13–24 months ago	43.1 (130)	45.0 (109)	44.8 (863)	44.6 (1102)		
Driving education	"dual"	19.0 (116)	20.0 (75)		19.4 (191)		
	"classic"	50.9 (116)	44.0 (75)		48.2 (191)		
	"L17"	30.2 (116)	36.0 (75)		32.4 (191)		
Driving frequency	Daily	73.8 (130)	45.9 (109)	57.2 (862)	58.0 (1101)	0.12	0.0015
	Weekly	24.6 (130)	51.4 (109)	36.8 (862)	36.8 (1101)		
	Rarely	1.5 (130)	2.8 (109)	6.0 (862)	5.2 (1101)		
Traffic fine	yes	53.9 (130)	31.2 (109)	40.4 (854)	41.1 (1093)	0.09	0.0024
	No	46.1 (130)	68.8 (109)	59.6 (854)	58.9 (1093)		
Recent traffic fine	yes	29.3 (116)	21.3 (75)		26.2 (191)		
	No	70.7 (116)	78.7 (75)		73.8 (191)		
Privacy attitude	very important	56.9 (130)	56.9 (109)	63.5 (854)	62.0 (1093)	0.04	0.17
	Not very important	43.1 (130)	43.1 (109)	36.5 (854)	38.0 (1093)		
Privacy settings	standard	20.0 (130)	17.4 (109)	21.7 (854)	21.0 (1093)	0.02	0.80
	Liberal	13.1 (130)	11.9 (109)	14.8 (854)	14.3 (1093)		
	Restrictive	66.9 (130)	70.6 (109)	63.6 (854)	64.7 (1093)		
Risk preferences	risk-averse	60.2 (113)			60.2 (113)		
	Risk-loving	39.8 (113)			39.8 (113)		
Car-centricity	car-centric	50.4 (113)			50.4 (113)		
	Not car-centric	49.6 (113)			49.6 (113)		

Although no app-based feedback was provided to the participants during the pre- and post-measurement phase, the scores were computed and stored in the background, and were used in the analyses presented in this paper. All scores were reset at the end of pre-measurement, such that at the start of the treatment phase participants had no information on their performance during pre-measurement. Moreover, we recorded information on how often participants engaged with the app and looked up their scores.

At the start of the pre-measurement, we communicated to all participants that at the end of the post-measurement a lottery with prizes amounting to 5000 Euro overall would take place. At the end of the pre-measurement, we then assigned participants randomly to one out of two treatment groups, which allowed us to test whether monetary incentives improve safety-relevant driving behavior:

- **"Feedback" (FB)**: chance of winning in the lottery was *independent* of (safety-relevant) driving skills,
- **"Feedback & Incentive" (FB & INC)**: chance of winning in the lottery was *dependent* on (safety-relevant) driving skills: the higher the average trip score at the end of the treatment phase, the higher the chances to win in the lottery; members of this group received weekly updates on their current chances of winning in the lottery, and their potential chances of winning if they had higher scores.[3]

[3] We used a random incentive mechanism to incentivize drivers by participating in a lottery. This addresses a potential stake size effect as it enabled us to pay a much higher amount to a subset of participants. On the other hand, incentivizing individuals through lotteries might be not as effective as providing fixed monetary rewards to all participants. Charness, Gneezy, and Halladay (2016) surveyed the experimental evidence for and against both options and concluded that the potential loss in motivation through participating in a lottery is likely to be minor and outweighed by the benefit of increasing the reward amount.

Out of the above mentioned 130 participants, 67 were assigned to the FB group, and 63 to the FB & INC group.[4] The design of the experiment can be considered a mixed 2 × 3 factorial design, with one factor group (FB vs. FB & INC) and three phases: pre-measurement (i.e., baseline), treatment, post-measurement. When conducting the lottery, draws were first taken for the FB & INC group according to the announced probabilities of winning. The remaining part of the 5000 Euro payout was then raffled among members of the FB group. No deception of any of the participants was involved here, since we had communicated to potential participants that there would be an overall payout of 5000 Euro. Payments were made via bank transfer.

2.3. Material

2.3.1. Smartphone app

All participants had to install the customized app (*eMentoring app*) on their smartphone for the duration of the field experiment, with platform-specific versions for Android and iOS, respectively. The app was customized for the purpose of this study by the firm Dolphin Technologies (Website: https://www.dolph.in/en/), but utilized indicators and measurements that have been developed and successfully used by the same company in commercial insurance telematics before.[5] An updated version of the app that has been used in this study can be downloaded under https://www.getgosmart.com/.

2.3.1.1. Scores. The safety-relevant driving behavior of the participants was measured along six different dimensions, based on data generated by smartphone sensors and then processed using dedicated algorithms. For each of the six dimensions, scores ranging from 0 (very unsafe) to 100 (very safe driving behavior) were computed. They are listed below, with the relative weights in the trip-specific overall score provided in brackets:

1. **Phone use (33%).** This indicator reflects distraction from using the smartphone during the car ride, by touching the screen or accepting phone calls (unless a hands-free car kit is used). The imposed penalties for using the phone while driving are evaluated relative to the number of theoretically feasible points, which increase in the distance and time driven. In the standard mode, the app is in the foreground showing a black screen with the note "vehicle in motion" plus a pictogram. Smartphones with an Android-based operating system automatically recognize the start of a car ride, bringing the app to the foreground without further user intervention being required. In contrast, on iOS-based smartphones, the app requires the user to actively take the app to the foreground at the start of a trip, otherwise a score of 0 for smartphone usage is assigned. This act is not difficult in itself, but might easily be forgotten.
2. **Speeding (20%).**This indicator reflects instances in which the local speed limit has been exceeded. It is based on GPS information collected every 2 s and sent to the receiving servers for further calculations. After comparing position data with map and speed limit data from OpenStreetMap (Website: https://www.openstreetmap.org/), two different penalties that lead to deductions from the maximum score of 100 may apply. Penalty (1) is based on the maximum deviation from the applicable speed limit. Any speed up to 100% of the applicable speed limit (e.g. 100 km/h on country roads) leads to a 0 point deduction. Driving double the applicable speed limit (or a transgression of the speed limit by more than 50 km/h) leads to a deduction by 75 points from the initial 100 points. The deduction due to any transgression of the speed limit in between is derived linearly. Penalty (2) is based on the average deviation from the applicable speed limit. The reference value is the (weighted) average deviation from the applicable speed limit in instances where the speed limit was exceeded. The deductions are then computed in the same way as for Penalty (1). If, after the second penalty, the value is less than 0, the score for speed is set to 0.
3. **Speeding in dangerous areas (17%).**This indicator reflects speeding (relative to the local speed limit) in so-called "dangerous areas" (e.g. vicinity of nursing homes, child care facilities, hospitals, schools). It is computed in the same way as the speed score, but the maximum deviation is designed more strictly and higher deduction values are used for calculation.
4. **Cornering (10%).**Identifies and evaluates singular cornering events. A comprehensive model of concrete physical sensing and classification has been developed to identify such events. The model has nine stages and has been optimized by machine learning algorithms. It is a protected business secret of Dolphin Technologies. Similar to the phone use and speed scores, deductions for cornering are evaluated relative to the number of theoretically feasible points, which increase in distance and travel time.
5. **Acceleration (10%).**Identifies and evaluates singular acceleration events in a similar way as cornering events.
6. **Braking (10%).**Identifies and evaluates singular braking events in a similar way as cornering and acceleration events. It can be shown that scores for cornering, acceleration, and braking are highly correlated, with correlation coefficients exceeding 0.8.

[4] Subjects who registered late, at the end of the pre-measurement period or during the treatment phase, were still allowed to participate, but they were assigned to a separate (control) group. Members of the control group did not receive any app-based feedback even during the treatment phase, but were allowed to participate in the lottery (with their chance of winning being independent of their driving skills). Members of that group were not included in the analyses presented in this study, as this group included only 32 individuals, most of whom dropped out of the experiment prematurely.

[5] Until the main analyses were completed, none of the authors had an affiliation with the developers of the app. However, in 2018, i.e. one year after the experiment took place, one author (K. Sallinger) became employed by Dolphin Technologies. Before, she was employed at the Austrian Board for Traffic Safety (Kuratorium für Verkehrssicherheit: KFV).

As stated above, the highest relative weight in the overall score is attached to the score associated with smartphone usage (33%). Inattention and distraction are the dominant cause of car crashes in Austria; they play a role in 37.6% of all car crashes (Statistik Austria, 2016). The target group of the study tends to be very smartphone-affine and hence distraction by smartphones is likely to be even more relevant in the context of this study. A number of studies (see for instance Caird, Johnston, Willness, Asbridge, & Steel (2014) for a meta-analysis on the effects of texting while driving) has shown the detrimental effect of smartphone usage on traffic safety. Moreover, phone use while driving is typically regulated legally: in Austria, phone use while driving is only legal if hands-free equipment is used.

Also the association between speeding behavior (i.e. violation of speed limits) and accident risk is well established in the relevant literature (e.g. Cooper, 1997; Shawky, Al-Badi, Sahnoon, & Al-Harthi, 2017), and regulated by Austrian federal and provincial laws: depending on how much and where the speed limit is exceeded, the penalties range from minor monetary fines to driver's license revocation and criminal proceedings. Accordingly, the scores for speeding and speeding in dangerous areas are weighted heavily in the overall score: together, they account for 37% of the overall score.

Cornering, acceleration, and braking account for only 10% each, as clearly specified legal rules do not apply to them, and drivers presumably have less control over them. For instance, sudden braking instances may be required due to external factors rather than actions associated with unsafe driving behavior. Moreover, perceptions, for example of very light to very hard cornering, acceleration, or braking, may be subject to a socio-cultural evaluation, which can differ widely (e.g. Özkan, Lajunen, Chliaoutakis, Parker, & Summala, 2006): what is considered strong acceleration by one driver, might be perceived as mediocre acceleration by another. This is in contrast to the speeding and phone use indicators, which can be objectively measured, independent of contextual factors, not at least because in most countries clear legal rules apply to both. For the same reasons, in the analyses, besides the overall score, we focused on the phone and the speed score. Note that it is not advisable to adjust the weights of the sub-scores in the overall score after the field experiment has taken place, as this is how the overall score had been presented to the participants. Their reaction to the overall score was thus with respect to the chosen weighting scheme.

Fig. 1 shows several screenshots of the *eMentoring app* in order to demonstrate how the feedback was provided to the participants of the study. Fig. 1a provides the aggregate overall score, which is the average of the trip-specific overall scores (weighted by their length) that are associated with all trips made by that specific participant since the start of the treatment phase (the scores during the pre-measurement were not visible to drivers and were not taken into account in the computation of the overall score). The screen also shows the overall number of trips as well as the overall distance (in km) and the overall duration (in hours) associated with all trips since the start of the treatment phase. Participants could filter the overall score, the number of trips, and the overall duration on a calendar basis (month, week, day).

Fig. 1b provides an example of the logbook functionality of the app, which allows the user to zoom into a single trip and see all six sub-scores associated with that trip. For each of the sub-scores, the user can further obtain a map-based overview (see Fig. 1c), in which the geographical location of specific events and their intensity (exceeding the speed limit, acceleration/braking/cornering events etc.) is shown.

2.3.1.2. App use & app use intensity. We were able to observe the extent to which a specific participant made use of the app during the treatment phase. This information did not exist during pre- and post-measurement, in which participants did not receive any feedback neither on past nor current trips. Specifically, we recorded (as a dummy variable) if the participant opened the app between two consecutive car trips. The corresponding person-specific *app use indicator* is then defined as the relative share of trips after which the app was opened. The indicator is hence defined on $[0, 1]$, implying that for instance a value of 0.8 means that the user opened the app after 80% of his/her trips.

In order to capture more specific interaction patterns with the app, we also define the (again person-specific) *app use intensity indicator*, which is the relative share of trips after which the app was opened to obtain information on a specific trip (presumably, the most recent trip). When opening the app, a user received feedback on his/her overall driving behavior through the overall score (visible on the start screen). However, if a user chose to obtain feedback on individual trips, (s) he was required to actively navigate through the app. The indicator for app use intensity is also defined on $[0, 1]$. A value of 0.8, for instance, implies that the driver had actively sought feedback on at least one individual trip after 80% of his/her trips.

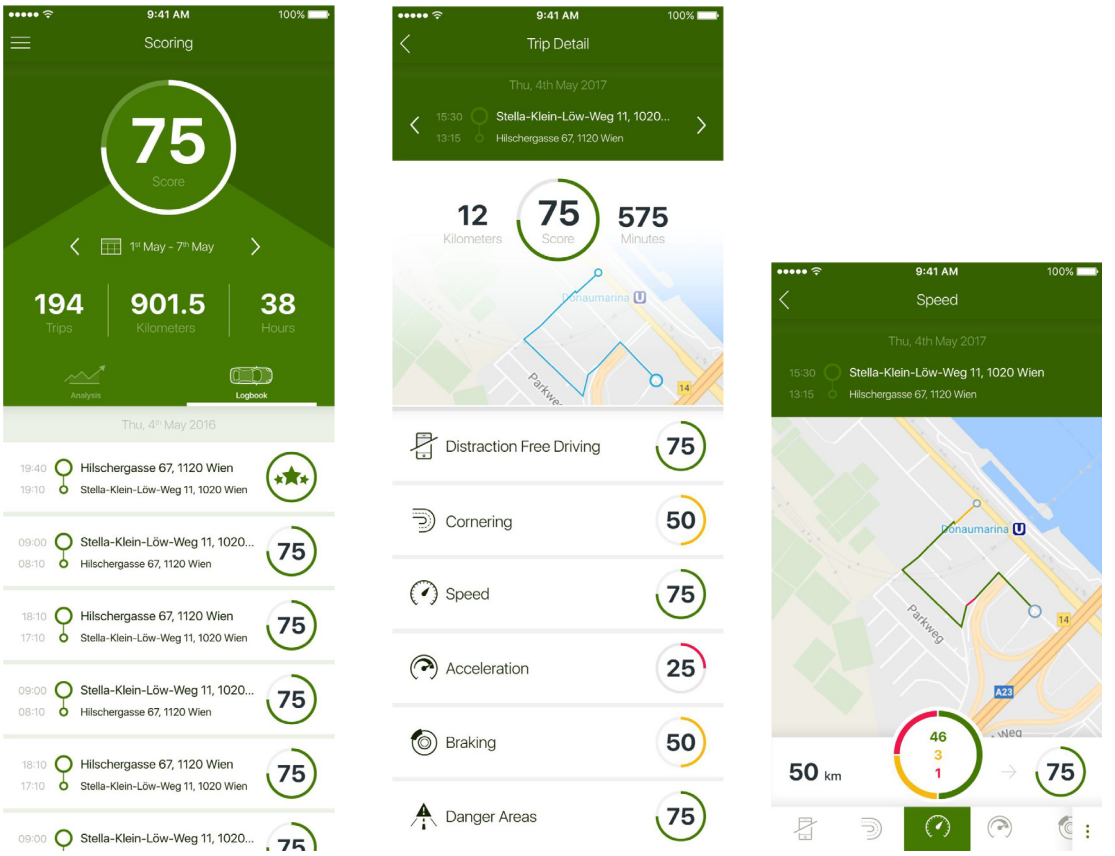
2.3.2. Dongle

The dongle had to be installed in the participants' cars in order to identify car trips relevant to the field experiment (the dongle connects via Bluetooth to the *eMentoring* app), but no further data was used from the dongle. The device is also capable of automatic accident detection and reporting, and functions as an emergency button and car-finder (see www.hierbox.com for more information). It is worth approximately 50 Euros, and participants were allowed to keep it and use those functions after the end of the field experiment.

2.3.3. Questionnaires

Two (obligatory) questionnaires had to be filled in by participants at the beginning and the end of the field experiment, respectively:

Questionnaire 1 (Q1). The initial questionnaire contained question items concerning socio-economic variables, driving behavior (e.g. past traffic fines), and the extent of driving experience. We also collected data on a large number of attitudinal



(a) Overview past trips (b) Overview specific trip (c) Indicator-specific overview for a specific trip (here: speed)

Fig. 1. Screens *eMentoring app*.

items, mostly related to traffic situations (whether respondents enjoy driving/ enjoy driving fast/ tend to stick to traffic regulations/ like driving long distances/ are nervous during lane change/ cars mean more to them than a pure means of transport etc.), but also concerning, technology affinity, and privacy concerns. Filling in the initial questionnaire was a prerequisite for participating in the field experiment.

Questionnaire 2 (Q2). Participants had to fill in a second questionnaire towards the end of the field experiment in order to be eligible to participate in the lottery. Besides questions focusing on the evaluation of the experiment, the second questionnaire contained questions that allowed us to determine risk preferences and "car-centricity" of the participants. Besides those participants who were actively participating in the field experiment until its end, we were able to motivate (through phone calls) some of the dropouts to participate in the second questionnaire.

2.3.4. Variables of interest elicited from the questionnaires

The following variables of interest were elicited from the above mentioned two questionnaires (the corresponding questionnaire number is stated in brackets). Further variables contained in the questionnaire but not used in any of the analyses, are not described here.

Privacy attitude (Q1). On a scale from one to five, how important is data protection to you? "Very important" corresponds to a score of 5, "less important" to 2–5, "not important" to 1.

Privacy setting (Q1). How would you consider your privacy settings on social media platforms such as Facebook or Instagram? (1) Clearly more public than default, (2) Slightly more public than default, (3) I use default settings, (4) Slightly more private than default, (5) Clearly more private than default, (6) I do not know my privacy settings, (7) I do not use social media. We group those answers into "standard" (3,6), "liberal" (1,2), and "restrictive" (4,5,7).

Population of home municipality (Q1). The population of the home municipality serves as a proxy for whether the participant tended to drive in rural or more urban areas. For instance, Grüninger (2008) showed that for young drivers in Bavaria (which is similar to Austria in terms of geographical and socio-economic structure) the accident risk was higher in rural areas. The following answer categories regarding the population of the home municipality had been presented to respondents: (1) up to 2.000, (2) up to 5.000, (3) up to 20.000, (4) up to 50.000, (5) above 50.000, (6) Vienna.

Driving frequency (Q1). The question regarding driving frequency had the following answer categories: (1) almost daily, (2) multiple times a week, (3) once or twice a week, (4) once or twice a month, (5) more rarely. These categories have been summarized to "daily" (1), "weekly" (2 and 3), and "rarely" (4 and 5).

Traffic fine (Q1). Have you received a traffic fine so far? Reference period: since obtaining the driving license (max. 2 years).

Recent traffic fine (Q2). Have you received a traffic fine in the past four months?.

Car-centricity (Q2). The concept of car-centricity is based on Holte, 2012 and Holte, Klimmt, Baumann, and Geber (2014). Applying factor analysis to a number of attitudinal statements, they identified a group of drivers characterized by high car-affinity and a (negative-) aggressive emotional state, which tends to translate into an aggressive driving style and higher accident risk. They were able to show that this group has the highest accident involvement rates and an above-average number of traffic violations. To elicit car-centricity (defined as a binary variable) for the participants of our study, we used a reduced version of the approach suggested by Holte (2012), which is based on those 14 attitudinal items that showed the highest factor loads in the original study.[6]

Risk preferences (Q2). It has been shown that drivers with higher degrees of risk aversion have lower rates of accident involvement than more risk-loving individuals (e.g. Turner & McClure, 2003). Our indicator for risk aversion is based on a simplified version of the survey method developed in Holt and Laury (2002), in which respondents were shown a series of questions, in each of which they needed to decide between a lottery (with 2 possible payouts) and a fixed payment. While the lottery remained the same across questions, the fixed payment increased, starting from the lower payout of the lottery until matching the higher payout of the lottery. The relevant indicator is then after how many questions the respondent switched from choosing the lottery to choosing the fixed payment. The earlier (s) he switched, the more risk averse (s) he is. If (s) he switched within the first 4 questions, (s) he is considered "risk-averse" (in question 5, the expected value of the lottery equals the fixed amount).

2.4. Recruitment

The recruitment took place mainly via social media, using Facebook advertisements. Complementary measures included the distribution of posters and flyers at driving schools, schools, universities, student dormitories, and the Austrian military. The complementary measures mainly took place in the Eastern part of Austria (including Vienna). Both the digital as well as the analogue campaign were centered around the question "How good are your driving skills?" and the slogan "Be part of the largest Austrian study for road safety and get rewarded" (both translated from German). The campaign emphasized that it is easy to join the experiment, and that each participant receives a small gift (the "Hierbox") and has a chance to win monetary payouts with a total value of 5000 Euro.

Our social media campaign for the recruitment of participants resulted in 1.642 million ad impressions, 359,800 individuals reached, and resulted in 12,500 clicks on the study website (where access to the (rather lengthy) initial questionnaire was provided). The complementary measures (250 posters and 750 flyers) generated 3018 visits of the study website. Among those 15,518 who visited the website, 4195 started the questionnaire, out of which 1690 finished the questionnaire. In turn, out of these 1690, 1102 were in our target group, meaning that they fulfilled the three above-mentioned criteria based on the information provided in the questionnaire. These 1102 then received an invitation to download the app and an on-board diagnostic (OBD) dongle was sent to them via postal mail. Among those 1102 persons who received the OBD, 245 successfully linked the app with the OBD dongle, which had to be installed in the participant's car and was used to identify relevant trips (194 installed the app, but did not connect it to the OBD dongle). We have no information on why so many persons who received the OBD, did not link it to the app. Out of the 245 persons who installed and linked the app and the OBD successfully, we have registered trip observations from 239 persons. See A for a diagrammatic exposition of the recruitment process (as well as sample definitions).

2.5. Procedures

The experiment took place in Austria in spring 2017. The pre-measurement phase lasted from March 1 to April 8, the treatment phase from April 9 to June 10, and the post-measurement phase from June 11 to July 9. The first questionnaire (Q1) had to be filled in before starting to participate, and the second questionnaire (Q2) became available at the end of the treatment phase. All participants took part in the field experiment voluntarily. During the recruitment process as well as when installing the app, all participants had been clearly informed about the data recorded, the data processing during and after the experiment, and they had to actively agreed to the terms of participation.

Those persons who registered late (at the end of the pre-measurement period or during the treatment phase) were still allowed to participate, but they got assigned to a separate (control) group, who did not receive any app-based feedback even during the treatment phase, but who were allowed to participate in the lottery (with their chance of winning being independent of their driving skills). Members of that group were not included in the analyses presented in this study for

[6] The reduced version had been developed in cooperation with the German Federal Highway Research Institute bASt (www.bast.de), which also published the original studies by Holte, 2012 and Holte et al. (2014), and MTO (www.mto.de), which is an advisory firm specialized in psychological research.

the following reasons: they did not have measurements during the pre-measurement phase (hence lacking comparable baseline measurement), and the group included only 32 individuals, most of whom dropped out of the field experiment prematurely.

2.6. Statistical methods

Test statistics We applied χ^2 or Fisher-exact tests to test for differences between different groups of individuals, which are further defined and described in the following sub-sections. Moreover, our evidence concerning self-selection biases (Table 1) and attrition biases (Table 2) stems from these tests.

Regression analyses We present the results of five standard ordinary least squares (OLS) regressions (all presented in Table 3). We also applied alternative, more complex estimation procedures, in particular linear mixed effects models, but they yielded results that are qualitatively and quantitatively similar to those derived from standard OLS models. OLS regressions, as opposed to statistical tests of differences (see above paragraph), are able to control for the influence of confounding variables.

1. We first explain the *aggregate overall score at the end of the pre-measurement* (hence before the participants received any app-based feedback) by means of socio-economic, attitudinal, and driving-behavior-related variables for those participants who completed both questionnaires (see first results column of Table 3). Given the fairly low number of

Table 2
Descriptive statistics on active participants vs. dropouts.

		FB		FB & INC		all	
Variable		$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.
App use	Active	0.61	0.08	0.60	0.11	0.61	0.10
	Dropout	0.62	0.19	0.61	0.17	0.62	0.18
App use intensity	Active	0.06	0.11	0.11	0.11	0.09	0.11
	Dropout	0.11	0.17	0.11	0.12	0.11	0.15
Av. # trips pre-measurement	Active	54.59	31.88	48.07	25.98	50.59	28.22
	Dropout	33.30	25.53	35.06	23.23	34.04	24.47
Av. # trips treatment	Active	92.18	47.32	77.82	34.97	83.36	40.28
	Dropout	37.42	37.84	33.53	30.34	35.79	34.76
Av. # trips post-measurement	Active	5.35	7.52	8.67	11.72	7.39	10.33
	Dropout	2.52	9.52	2.42	8.15	2.48	8.92
Average trip length (in km)	Active	18.57	9.60	17.30	11.80	17.79	10.90
	Dropout	20.55	18.66	17.14	11.40	19.12	16.04
Overall score at the end of							
Pre-measurement	Active	58.82	11.54	64.70	12.29	62.43	12.22
	Dropout	61.86	15.05	62.78	13.90	62.24	14.50
Treatment	Active	60.41	12.75	70.52	13.36	66.61	13.90
	Dropout	63.04	15.18	65.61	13.69	64.12	14.54
Post-measurement	Active	59.94	12.21	70.22	13.13	66.25	13.61
	Dropout	62.90	15.25	65.58	13.69	64.02	14.60
Δ Treatment - Pre-measurement	Active	1.59	8.28	5.82	4.76	4.18	6.60
	Dropout	1.18	9.78	2.83	9.04	1.87	9.46
Speed score at the end of							
Pre-measurement	Active	57.13	18.66	58.96	20.41	58.25	19.55
	Dropout	60.25	16.20	55.04	13.85	58.07	15.39
Treatment	Active	57.36	18.33	62.69	17.45	60.63	17.78
	Dropout	59.99	14.83	55.41	12.69	58.07	14.08
Post-measurement	Active	57.36	18.31	62.57	17.51	60.56	17.80
	Dropout	59.98	14.83	55.38	12.69	58.06	14.09
Δ Treatment - Pre-measurement	Active	0.23	4.16	3.73	6.59	2.38	5.97
	Dropout	−0.26	4.74	0.37	5.42	0.003	5.01
Distraction-free driving score at the end of							
Pre-measurement	Active	33.41	31.93	60.56	34.13	50.07	35.53
	Dropout	42.58	34.58	54.58	34.00	47.60	34.66
Treatment	Active	34.65	32.93	63.82	33.22	52.55	35.74
	Dropout	42.47	34.24	54.63	33.17	47.56	34.14
Post-measurement	Active	34.60	32.92	63.83	33.24	52.53	35.76
	Dropout	42.44	34.22	54.64	33.18	47.55	34.13
Δ Treatment – Pre-measurement	Active	1.24	15.74	3.27	10.51	2.48	12.65
	Dropout	−0.11	7.74	0.05	8.18	−0.04	7.88
Sample size	Active (n_{01})	17		27		44	
	Dropout (n_{02})	50		36		86	
	Overall	67		63		130	

Table 3
Determinants of the aggregate overall score at the end of the pre-measurement (1), of the app use intensity (2), and differences in the overall/speed/phone use score between end measurement and end treatment phase (3–5). Note: "Average trip length (in km)" refers to the pre-measurement for regression (1) and to pre-measurement and treatment phase in regressions (2)–(5).

			Δ Treatment - Pre-Measurement		
	Score at the end of Pre-Measurement (1)	App use intensity (2)	Overall (3)	Speeding (4)	Distraction-free Driving (5)
Group: "FB & INC"	2.378	1.567	3.628**	1.373	2.972
	(2.297)	(2.906)	(1.596)	(1.158)	(2.152)
App use intensity			0.220**	0.214***	0.056
			(0.086)	(0.062)	(0.116)
App use intensity * av. trip length (treatment; in km)			−0.001	−0.002*	−0.001
			(0.001)	(0.001)	(0.002)
Score pre-measurement)55,70] - mid tercile		3.470	−4.945**	−1.842	−5.506**
		(3.532)	(1.943)	(1.410)	(2.620)
Score pre-measurement)70,91] - high tercile		8.597*	−9.780***	−2.636	−9.042***
		(4.385)	(2.454)	(1.780)	(3.308)
Smartphone platform: iOS	−6.289***	−5.745*	−3.100*	2.569*	−3.942
	(2.364)	(3.192)	(1.785)	(1.295)	(2.406)
Average trip length (in km)	−0.308***	0.323***	−0.087	−0.031	0.014
	(0.099)	(0.098)	(0.073)	(0.053)	(0.098)
Number of Trips (treatment phase)		−0.074**	−0.003	0.016	−0.024
		(0.033)	(0.019)	(0.014)	(0.025)
(Reported) driving frequency: more than weekly	−0.763	1.572	2.813	0.319	2.229
	(2.531)	(3.206)	(1.765)	(1.280)	(2.379)
Age >= 20 years	7.303**	−0.999	1.310	3.031**	2.250
	(2.955)	(3.654)	(2.018)	(1.464)	(2.720)
Gender: female	6.097**	−0.695	−0.619	0.712	1.965
	(2.462)	(3.270)	(1.790)	(1.298)	(2.412)
Education: high school exam or higher	−3.022	4.238	2.708	−0.749	3.526
	(2.971)	(3.726)	(2.054)	(1.490)	(2.769)
Net income > 2500 Euro	−3.728	−2.147	0.314	1.802	0.775
	(2.254)	(2.908)	(1.601)	(1.161)	(2.158)
No recent traffic fine	7.773***	−2.498	−2.467	−0.239	−1.334
	(2.499)	(3.323)	(1.824)	(1.323)	(2.459)
Not car-centric	3.760	1.456	3.990**	0.309	2.864
	(2.622)	(3.351)	(1.836)	(1.332)	(2.476)
Constant	60.603***	9.079*	4.773	−3.381	2.878
	(3.743)	(5.152)	(3.141)	(2.278)	(4.234)
Observations	99	99	99	99	99
R^2	0.401	0.317	0.431	0.280	0.239
Adjusted R^2	0.333	0.212	0.328	0.150	0.102
Residual Std. Error	10.754 (df = 88)	13.357 (df = 85)	7.306 (df = 83)	5.299 (df = 83)	9.849 (df = 83)
F Statistic	5.892*** (df = 10; 88)	3.034*** (df = 13; 85)	4.184*** (df = 15; 83)	2.149** (df = 15; 83)	1.741* (df = 15; 83)

Note: *p<0.1; **p<0.05; ***p<0.01.

observations (99), the regression serves mainly the purpose of understanding the determinants of the score in the context of our field experiment.

2. In a further OLS regression analysis, we explain which variables determined the share of trips after which the *app has been used "intensively"* (see second results column of Table 3).
3. to 5. We explain which variables determine the *difference between the aggregate (overall/speed/phone) score at the end of the treatment phase and at the end of the pre-measurement* (see third to fifth results columns of Table 3). These differences (if positive) can be interpreted as an indicator for learning based on the feedback provided by the app. Note that the overall score includes all six score categories according to the weights presented in Section 2.3.1.

3. Results

This section contains the main results derived in this paper. It starts by providing definitions and descriptions of multiple groups (participants vs. non-participants, active participants vs. dropouts, sample used in the regressions) including evidence on participation and attrition biases, and an introduction to the corresponding behavioral data (Section 3.1). It then

proceeds with presenting our evidence regarding the effects of feedback and incentives on safety-relevant driving scores (Section 3.2), determinants and effects of the app use intensity (Section 3.3), the role of the phone platform (Section 3.4), and the role of socio-economic and attitudinal variables (Section 3.5).

3.1. Data analysis

In this part of the paper, we first give an overview of the summary statistics of participants and non-participants in terms of socio-economic characteristics as well as travel-related indicators and preferences, providing evidence on self-selection biases(Section 3.1.1). Section 3.1.2 defines and compares active participants and dropouts along behavioral indicators, providing evidence on attrition biases. Finally, Section 3.1.3 defines the sample that is used in the subsequent analyses. A diagrammatic exposition of the interrelations between these sub-samples is provided in A.

3.1.1. Participants vs. non-participants: self-selection

We distinguish between participants considered in our further analyses, participants who were not considered in our further analyses, and non-participants.

Participants considered in the analyses $\mathbf{n_0} = 130$. In our analyses, we only considered subjects with at least 10 trips recorded during pre-measurement. In order to quantify a learning effect, we need to calibrate the user-specific driving performance during the pre-measurement phase. Calibration is only meaningful if it is based on a minimum number of trips, which has been derived from a sensitivity analysis and thereby set to 10. By that, mainly users who registered late in the pre-measurement phase had been excluded from the analyses. Below the number of 8 to 10 trips, the calibrated driving behavior during the pre-measurement exhibited random fluctuations for most users, but was fairly stable above 10 trips.

Participants not considered in the analyses $\mathbf{n_1} = 109$. Those subjects with less than 10 trips recorded during pre-measurement either drove only infrequently during the pre-measurement, or registered fairly late, such that they could not reach 10 trips before the start of the treatment phase. As a consequence, we only had a poor indicator for this group's baseline behavior. We, however, refrained from applying a selection criterion to behavior during the treatment phase (such as a minimum number of recorded trips during the treatment phase), as the criterion might be correlated with group membership (FB vs. FB & INC), potentially reducing or augmenting the differences between the two groups, and hence causing biases in the results.

Non-participants $\mathbf{n_2} = 863$. Non-participants are subjects who filled in the initial survey and fulfilled the criteria for being in the target group, but who decided not to participate in the field experiment.

Descriptive statistics for these three groups are presented in Table 1. Note that some variables in Table 1 are only available for participants (e.g. smartphone platform), and some variables only for those participants who filled in the second questionnaire (e.g. risk aversion, car centricity). Moreover, the number of persons for whom information on a specific variable is available varies due to the fact that not all respondents answered all question items in the two questionnaires, and that not all of those classified as participants filled in the second questionnaire.

Self-selection We found that participants (with at least 10 trips during the pre-measurement) and non-participants are very similar in terms of the investigated variables (see Table 1). The most distinct difference is in the reported driving frequency, where a much larger share of those with at least 10 trips during the pre-measurement indicated to drive daily (73.8%) compared to non-participants (57.2%). Secondly, we found a difference regarding the share of persons who had received a traffic fine: it is higher among participants (53.9 vs. 40.4% among non-participants). Finally, attitudes towards privacy as well as the chosen privacy settings (on social media platforms, smartphones etc.) do not differ significantly. However, it is worth noting that the relative percentage of those who consider privacy as (very) important is 6.6 percentage points higher among non-participants.

3.1.2. Active participants vs. dropouts: attrition

We further split the participants with at least 10 trips during the pre-measurement (n_0 = 130, see Table 1) into active participants and dropouts:

Active participants $\mathbf{n_{01}} = 44$. Active participants were defined as those who had checked the app after more than 40% of the trips over the course of the treatment phase and for whom we still registered trips during the post-measurement. The 40%-threshold corresponds to the lowest decile in terms of the app use indicator. It led to an elimination of 13 drivers, 3 of which did not have any trips during post-measurement and hence would also be classified as dropouts if they had used the app more. 6 of the eliminated drivers had many days between two trips. Only the remaining four drivers drove fairly regularly (including during the post-measurement). We nevertheless classified them as dropouts, as the learning effect is expected to realize only with regular usage of the app.

Dropouts $\mathbf{n_{02}} = 86$. Dropouts were those who did not fulfill the condition of being an active participant, either because they checked the app at most after 40% of the trips during the treatment phase or they were not active during post-measurement (or both).

Among those participants with at least 10 trips during the pre-measurement, the attrition rate hence amounts to 66%. Especially the inactivity in the post-measurement disqualified many participants to be labelled "active participant", also implying that we could not make strong inferences from the data collected during post-measurement due to a lack of

observations. Table 2 provides information on the performance regarding scores and app use of the active participants and the dropouts, also separately for the FB and the FB & INC group.

Attrition. Participants could drop out of the study either by actively de-installing the app or by hardly ever engaging with the app (see Fig. 2f). Table 2 provides evidence for differences between participants who were active throughout the entire field experiment including the post-measurement and dropouts (who either had no observations during the post-measurement or did not check their app regularly; see Section 3.1.2). Among dropouts, we observed a lower trip frequency during the pre-measurement, the treatment phase, and the post-measurement. In terms of overall, speed and phone use score, we found that active participants and dropouts had comparable aggregate scores at the end of the pre-measurement; at the end of the treatment phase, the group of active participants and the FB & INC group exhibited a higher improvement in scores compared to the dropout group and the FB group, respectively: for the overall score, the improvement amounts to 6.7% for the group of active participants, compared to 3.0% for the dropouts. We found no significant differences between active participants and dropouts concerning other (socioeconomic, attitudinal, mobility-related) variables (similar to those listed in Table 1).

We could further infer some reasons for prematurely dropping out from the second questionnaire. Especially, the app functionality seemed to play an important role. First, some participants found the scoring rather intractable and counterintuitive, especially in the iOS app and in specific situations such as tunnels. Second, some complained about wrongful deductions in the speed score, possibly due to an incorrect mapping of the trip to the road network, or due to speed limits being incorrectly registered in OpenStreetMap (which is used by the app to determine speed limits). Third, multiple respondents commented that the driving style suggested by the app is in general too conservative and hinders traffic if implemented as such. Finally, there were complaints about the app's high power usage.

3.1.3. Defining the sample used in regressions

In the models presented in this section (Table 3), we included those participants who had at least 10 trips during the pre-measurement and who used the app at least after 40% of their trips. Out of these 117, we only took into account those 99 who had completed both questionnaires.

3.2. The effects of feedback and incentives

Participants had been randomly assigned to the FB and the FB & INC group, implying that the two groups (at the beginning of the pre-measurement), as expected, did not differ significantly in key variables including socio-economic variables, phone platform, trips lengths or traffic fines. We moreover did not observe a statistically significant trend in the scores during the pre-measurement phase. This is an indication that scores did not improve in the absence of feedback and/or incentives. Section 3.2.1 presents descriptive evidence on the effects of feedback and incentives on safety-relevant driving behavior, whereas Section 3.2.2 presents regression-based evidence (controlling for confounding variables).

3.2.1. Descriptive evidence

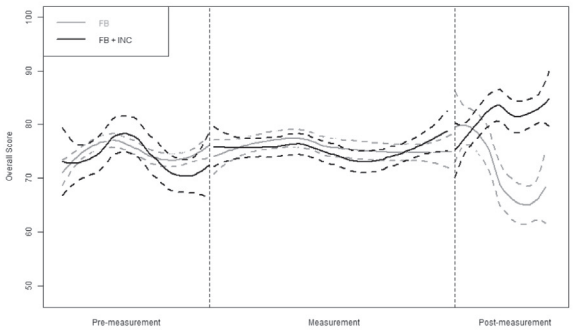
First, graphical evidence for the difference between the "Feedback (FB)" and the "Feedback and Incentive (FB & INC)" group is presented (see Fig. 2). It shows how the overall, phone and speed scores, the number of drivers that were active on a specific day, as well as the app use indicators evolved over the duration of the field experiment. The figures displaying the scores (Figs. 2e) and the figure showing the number of drivers (Fig. 2f) cover the entire field experiment including pre- and post-measurement; the figures displaying the app-use indicators (Figs. 2h) cover only the treatment phase, as during the pre- and post-measurement phases no feedback was provided. They include all participants with at least 10 trips during pre-measurement.

As shown in Fig. 2f, the number of active drivers (i.e. those conducting at least one trip) on a given day varied over time. Due to the lower number of drivers at the beginning and at the end of the study, the numbers shown in the figures for those time periods have little significance. Evidence from Table 2 shows that for active participants, the scores at the end of the post-measurement did not differ significantly from those at the end of the treatment phase. This holds for both the FB as well as the FB & INC group. However, the number of observations was fairly small.

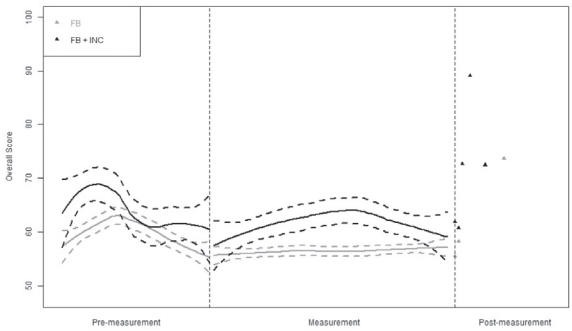
Fig. 2 presents the overall and the phone use score separately for iOS and Android, since the phone use score diverged significantly between the two platforms as shown by Figs. 2c (Android) and 2d (iOS) due to differences in the functionality as described in Section 2.3.1. As the phone use score makes up one third of the overall score, also the overall score differs substantially between the platforms (albeit less). We further found a small, but visible upward trend for the speed score for the FB & INC group (Fig. 1c) over the course of the treatment phase. In the remaining score plots, the pattern is less distinct.[7]

Table 2, which comprises the group comparisons between members of the FB and the FB & INC group, as well as between active participants and dropouts, indicates that at the end of the treatment phase, the FB & INC group exhibited a higher

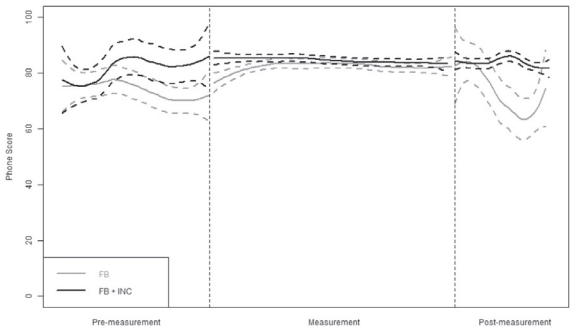
[7] Although participants had been assigned randomly to the FB and the FB & INC group at the end of the pre-measurement, Figs. 2c and 2d covering the phone use scores might give the impression that the FB & INC group performed better already in the pre-measurement. However, this pattern had been generated randomly. Rather, it is the consequence of averaging only over few observations (the number of group-specific observations shown in Fig. 2f, again sub-divided by platform) as well as the fairly binary distribution of the phone use score (0 vs. 100).



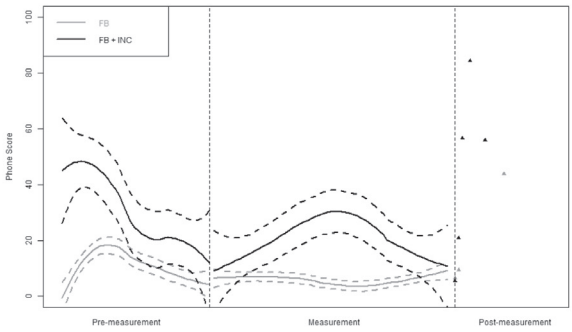
(a) Overall score (Android)



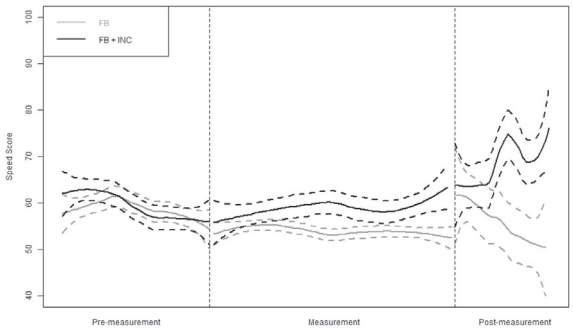
(b) Overall score (iOS)



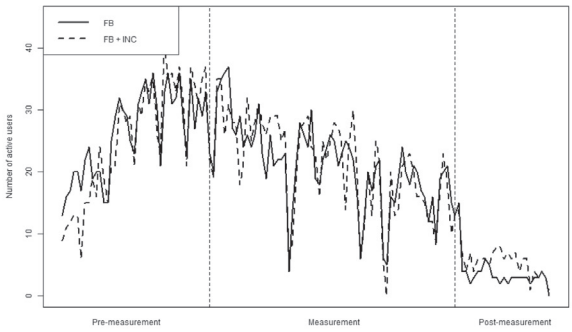
(c) Distraction-free driving score (Android)



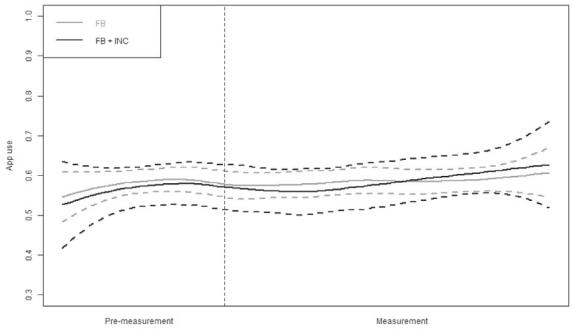
(d) Distraction-free driving score (iOS)



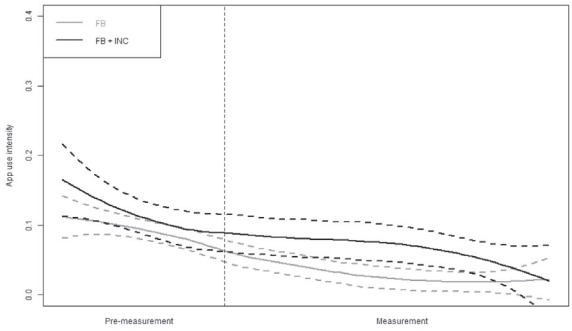
(e) Speeding score



(f) Daily number of drivers



(g) App use



(h) App use intensity

Fig. 2. Scores and usage patterns over time. The non-parametric Loess curves (and 95% confidence intervals) are based on the median score of those who made a trip on a specific day. Due to a lack of observations, the Loess curve is not shown for the post-measurement in the iOS-specific Figs. 2b and 2d. Timeline: start of the pre-measurement: 28 February 2017; start of the treatment phase: 9 April 2017; start of the post-measurement: 12 June 2017; end of the post-measurement: 8 July 2017.

improvement in scores compared to the FB group, respectively: for the overall score, the improvement amounts to 9.0% for the FB & INC group (p-value from two-sided t-test: 0.10). According to the descriptive analysis, the improvement of 2.7% for the FB group is not significantly different from 0 (p-value: 0.70). Also improvements in speed and phone scores are not significantly different from 0. Finally, app use barely differed between the FB and the FB & INC group (Fig. 2g), whereas the app use intensity indicator seems to be slightly lower for the FB than the FB & INC group (Fig. 2h).

3.2.2. Regression-based evidence

As expected, we found no difference between the FB and the FB & INC group when the aggregate overall score at the end of the pre-measurement is explained by, among other variables, group membership (see first results column in Table 3). Over the course of the treatment phase, members of the FB group improved slightly in terms of the overall score, as indicated by the positive constant (significant at the 10% level) (see regression (3) of Table 3). The improvement in terms of the overall score of members of the FB & INC group was in turn significantly higher than the improvement of members of the FB group. Compared to the descriptive comparisons in Table 2 (as discussed in Section 3.2.1), the regression model thus shows stronger evidence for an improvement in overall scores for both groups, with the improvement being higher for the FB & INC group.

Regarding the speed and the phone use score, the coefficient estimates also point in the direction of the FB & INC group improving more over the course of the treatment phase than the FB group, however, the corresponding coefficient estimates were not significantly different from 0. Note that the estimated model with the difference in terms of the overall score as dependent variable have the best fit (adjusted R^2 of 0.39), compared to an adjusted R^2 of 0.21 for the difference in terms of the speed score, and an adjusted R^2 of 0.16 for the difference in terms of the phone use score.

3.3. App use intensity

3.3.1. Determinants of app use intensity

The second results column of Table 3 shows that drivers with a relatively high score in the pre-measurement tend to have a higher app intensity indicator. The number of trips during the treatment phase is negatively associated and the average trip length is positively associated with the app use intensity indicator. Group membership did not significantly affect the app use intensity: members of the FB group thus chose for a similar feedback intensity as members of the FB & INC group (see also Fig. 2h, where a slightly higher app use intensity for the FB & INC group is visible, but the difference turned out to be not statistically significant in the regression analysis). Fig. 2g further shows that the app use indicator was fairly stable over the course of the treatment period (around 0.55–0.6), whereas the app use intensity indicator exhibited a downward trend (i.e. the relative share after which feedback on a specific trip had been viewed). It declined from 0.1 to 0.15 to below 0.05 (Fig. 2h).

We tested in a separate (trip-based) model (results are available upon request) whether participants had a higher likelihood to interact with the app (in terms of the app use as well as the app use intensity indicator) after having a below-average trip score; however, this did not seem to be the case, which is consistent with the finding that below-average scores do not have a higher likelihood of being followed by an above-average score than a below-average score.

3.3.2. Effects of app use intensity on the overall, speed and phone use scores

We found that app use intensity is positively correlated with the overall and the speed score, but not with the phone use score (see regressions (3)–(5) in Table 3). Only for the speed score also the interaction between average trip length and the share of trips after which the app had been used intensively matters. Even though participants with higher scores in the pre-measurement checked the app more frequently (see Section 3.3.1 and the second results column in Table 3), their improvement in scores was significantly lower than for those with low scores during pre-measurement (the reference group are those with scores lower than 55). Although the app use intensity was chosen by the participants and hence not exogenous, we expect endogeneity to be of minor importance here, since participants were unaware of the dependent variable (i.e. the differences in the aggregate scores between treatment and pre-measurement); they had no knowledge of their scores during the pre-measurement and they only became aware of the aggregate improvement over the course of the treatment phase at the end of the treatment phase.

3.4. The role of the phone platform

3.4.1. Overall score at the end of the pre-measurement

Persons with iOS-based smartphones had significantly lower overall scores at the end of the pre-measurement (see first results column in Table 3). This is most likely due to the different functionality compared to Android-based smartphones as described in Section 2.3.1: iOS users had to bring the app to the foreground manually in order to receive the full points on the phone use score.

3.4.2. Treatment phase

Figs. 2c and 2d show that overall and phone use score for those with Android-based smartphones exhibited little change over the course of the treatment period. For the iOS-based users in the FB & INC group, we observed an increase in the phone

use score (translating into a similar picture for the overall score) in the first half of the treatment phase, followed by a decrease of similar size. Controlling for other variables of interest, the regressions presented in Table 3 show that having an iOS-based smartphone lowered improvements in the overall and phone use score at the end of the pre-measurement and at the end of the treatment phase, but had a positive effect on the speed score. These regressions also indicate that persons with iOS-based smartphones had a lower app use intensity.

3.5. The role of socio-economic and attitudinal variables

3.5.1. Overall score at the end of the pre-measurement

We found that females, participants above the age of 20, and participants who had not received a traffic fine recently had on average a higher overall score at the end of the pre-measurement. Drivers with higher average trip lengths tended to have a lower score. Other variables that were considered, but found to be nonsignificant (which might, however, be due to the small sample size) were education, population of the home municipality, risk aversion, car-centricity, and reported driving frequency.

3.5.2. Treatment phase

Socio-economic and attitudinal variables had only little explanatory power in all regressions regarding improvements in scores over the course of the treatment phase. Subjects who were not car-centric improved more in terms of the overall score than car-centric subjects. Also participants above the age of 20 tended to have a higher improvement in scores (in particular for speed).

4. Discussion

4.1. Results

Based on the regression results, we found that feedback had a positive (albeit small) effect on safety-relevant driving behavior, and that the effect was higher if feedback was combined with an incentive for safe driving. In line with the literature (Mullen et al., 2015; Reagan et al., 2013), these findings imply that extrinsic motivation in the form of an increased chance to win the lottery is effective in augmenting the learning effect. However, the effect sizes were small in our study and only became significant when controlling for confounding variables in a regression framework.

The effects also differed across score types. The improvements for the FB & INC group over the course of the treatment phase were significantly higher compared to the FB group only for the overall score (although the corresponding coefficients point in the same direction also in the speed and phone use score regressions). The significant effect in the overall score was thus probably the compound result of slight improvements in all sub-scores. This suggests that it is important to not only look at speed (as many earlier studies including Reagan et al. (2013) and Mullen et al. (2015) did), but also at other indicators for safety-relevant driving behavior.

We found few person-specific factors that had an influence on whether an individual benefits from the app in terms of improving his/her safety-relevant driving behavior. Older participants had a higher improvement in scores, presumably because they are more mature and willing to incorporate feedback in their driving-related decision making. Also subjects *not* classified as car-centric improved more. Car-centricity is a compound psychological construct, which includes a multitude of factors, some of which might be correlated to personality traits such as over-confidence, which in turn might render persons less likely to act upon feedback (Holte, 2012; Holte et al., 2014).

Generally, we found that those participants who chose to interact with the app more frequently (in terms of checking the results for specific trips), all other things being equal, exhibited stronger improvements in overall and speed scores over the course of the treatment phase. There was no effect of the app use intensity on the phone use score, which is plausible, as the computation of the phone use score was easy to understand and did not require "intense" usage of the app. We also found a significant downward trend with respect to app use intensity, possibly due to learning effects concerning the functionality of the app (the app use indicator, however, stayed roughly stable). Interestingly, we did not observe any difference in app use (intensity) between the FB and the FB & INC group, although the latter performed better in terms of scores.

We only had few observations for the post-measurement, which renders it difficult to infer what happened after feedback from the app was no longer available. We can only tell that the scores at the end of the post-measurement were not significantly different from those at the end of the treatment phase. This is an indication that learning effects persist, however, our data did not allow for conclusive results here.

We did not find much evidence on self-selection. Only few variables were significantly different between participants and non-participants. One was that individuals who had already received a traffic fine were overrepresented among participants; they might have had a higher interest in participating in a traffic safety experiment in order to avoid future fines. Privacy considerations seemed to play only a minor role. A likely reason why we observed little self-selection is that the group of non-participants consisted of those who filled in the first questionnaire (and fulfilled the criteria for being in the target group) and hence showed some interest in participating in the experiment, but ultimately decided against participating (see Section 3.1.1). Self-selection biases may have been more prominent if we had data also from those novice (Austrian) drivers who did not fill in the first questionnaire.

Attrition was more evident than self-selection: more than half of the participants dropped out over the course of the experiment. We found that there is a positive association between dropping out and low improvement in scores, which seems to be not only due to dropouts having had less overall exposure to the feedback; also the average improvement per trip registered during the treatment period was higher for active participants than for dropouts. Given the fairly equal app use rates across active participants and dropouts, which suggest that no substantial difference in terms of motivation levels was present, the lack of improvement in scores might have had contributed to the decision to quit the field experiment.

Various indications illustrate that our app measurements are plausible and valuable. (1) The correlation of the aggregate overall score during the pre-measurement and the aggregate overall score during the treatment phase amounts to 0.8 for participants. The strong correlation is an indication that the app indeed provided consistent feedback. (2) The results we derived when explaining the overall score at the end of the pre-measurement are widely in line with the literature. For instance, consistent with comparable studies (especially those based on a representative, larger sample, e.g. Sarma et al. (2013)), we found that females had significantly higher scores than males, possibly due to a more conservative, less risky driving style. (3) Persons who had received a traffic fine recently had lower scores, which is yet another indication that the app indeed reflects safety-relevant driving skills. (4) The model that explains changes in the overall score over the course of the treatment period had the highest explanatory power (as compared to the models explaining changes in the speed and phone use score). This suggests that the overall score was less affected by unobserved factors, and hence (as expected) is the most solid indicator for learning among the (sub-) scores.

Our findings also underline the role of technology. iOS users had to bring the app manually to the foreground in order to receive the full points on the phone use score, causing iOS users to have significantly lower points than Android users. Having an iOS-based smartphone had a significant negative impact on improvements in the overall and phone use score during the treatment phase, but a positive effect on the speed score. This might reflect a substitution effect: iOS users may have found it more challenging to improve their phone use score, and hence tried to improve relatively more in the other sub-scores (such as speed). Subjects using an iOS-based platform also had a lower app use intensity indicator, which also may have been a consequence of the more tedious app functionality, discouraging them from more intensive app use.

4.2. Limitations

While we did not find substantial differences between participants and non-participants, the sample of non-participants might not have been entirely representative of novice drivers in Austria. By filling in the questionnaire, the surveyed group of non-participants indirectly indicated an interest in participating in the field experiment, and are thus expected to be more similar to participants than those who did not fill in the questionnaire including (1) those who did not see the advertisement for the study (for instance, because they do not use Facebook or have only limited personal information on their Facebook account, making it difficult to identify them as members of the target group), (2) those who saw the advertisement but did not react to it (possibly due to a lack of interest and/or time), (3) those who went on the study website but did not fill in the survey (possibly because the text on the advertisement did not mention the necessary efforts involved in participating in the study as well as the relatively minor rewards, both of which were stated on the website), and (4) finally those who did not complete the survey (probably realizing the length and level of detail required when filling in the survey). Not having any data on non-participants who did not fill in the survey and their reasons behind not participating in the study limits our ability to draw conclusions on the general willingness of (Austrian) novice drivers to accept technology to promote road safety.

The relatively low number of 130 participants limits the feasibility of our empirical analysis and the interpretation of our results. The statistical insignificance of some of the coefficients in our regression analysis might have therefore been due to the lack of statistical power. Moreover, the field study lacked a proper control group of drivers who did not receive any feedback throughout the experiment. We thus could not properly control for a general learning effect over the course of the study and might have overestimated the learning effect due to app-based feedback. However, the absence of an upward trend in scores during the pre-measurement suggests that no improvements would have occurred without feedback (and incentives).

Due to the pronounced attrition at the end of the treatment phase, we also lacked statistical power to examine whether the improvement in safety-relevant driving skills was persistent or just temporary. Among the 44 participants that were active throughout the post-measurement phase, there is some empirical evidence for a persistent learning effect.

There are also limitations related to the smartphone app and the scoring. In our study, smartphones were used as a platform to measure safety-relevant driving. Paefgen et al. (2012) showed that these smartphone-based measures are less reliable compared to those measured via a fixed platform. More specifically, smartphone-based measures tend to overestimate critical driving events, such as braking. Moreover, before conducting the study, we fixed the relative weights of the six subscores to best reflect the relative importance for safe driving as reflected in the overall score. This is, however, not to claim that the choice of the weighting scheme for sub-scores is the only possibility nor that it is optimal. It was outside the scope of this study to test different weighting schemes. If we had applied relative weights, drivers might have reacted differently to the overall score and our results might have been somewhat different.

There are also some patterns in our results that may be artifacts of the underlying scoring algorithm. First, even though participants with higher scores in the pre-measurement checked the app more frequently (possibly because specific personality traits (e.g. diligence, prudence) affect both the driving behavior as well as the willingness to learn and to be a "well-behaved" participant in the field experiment), their improvement in scores was significantly lower than for those with

low scores during pre-measurement. This might be due to the underlying scoring algorithms, which allow for little improvement if a person had a safe driving style already. Second, we found that the average trip length is negatively correlated to improvements in the overall scores, which was probably an artifact of how the underlying scoring model weights singular events on a given trip.

Finally, incentivizing participants with fixed monetary rewards, instead of lotteries, might have led to different results. Charness et al. (2016) surveyed the experimental evidence and concluded that the potential loss in motivation through participating in a lottery is likely to be minor and outweighed by the benefit of increasing the reward amount. Incentivizing with a fixed but lower amount might thus have led to lower improvements in safety-relevant driving.

5. Conclusions

In this field experiment, we provided app-based feedback on safety-relevant driving behavior to novice drivers and measured the intensity at which drivers use the app. The field experiment was conducted among 130 drivers in Austria, and consisted of a pre-measurement, a treatment, and a post-measurement phase. In the treatment phase, drivers received feedback on six different indicators for safety-relevant driving skills (phone use, speeding, speeding in dangerous areas, cornering, acceleration, and braking). In the other two phases we recorded their driving behavior, but did not provide them with feedback. Participants were randomly assigned to one of two groups: one group only received the feedback, whereas the other group was additionally offered a monetary incentive for higher scores on the indicators for safety-relevant driving skills.

We found that feedback led to small improvements in safety-relevant driving skills, and that the improvements were larger but still moderate if monetary incentives for improving safety-relevant driving skills were present. Like earlier literature we therefore found limited evidence of intrinsic motivation leading to improvements in safety-relevant driving skills (Guttman & Gesser-Edelsburg, 2011; Musicant & Lotan, 2016). Our finding that participants who received monetary incentives increased their safety-relevant driving skills significantly more hints at the potential role of extrinsic motivation for improving traffic safety. Similar results have also been obtained by Reagan et al. (2013) and Mullen et al. (2015), however, the former study relied on real-time feedback once a driver exceeded the local speed limit, and the latter was based on a laboratory setup. Both focused only on speeding, and had no emphasis on young, novice drivers.

We also derived evidence on both the willingness of novice drivers to learn about their safety related driving skills and the extent of such learning. Our results show that drivers who drove relatively safely before receiving app-based feedback used the app more intensely, and that for a given score during pre-measurement, higher app use intensity led to larger improvements in driving skills. These results are largely in line with the literature on self-regulated (online) learning environments such as massive open online courses (MOOCs), in which participants can choose how much time and effort they (voluntarily) wanted to spend on a course (e.g. De Barba et al., 2016). Nevertheless, the improvements were largest for drivers who drove least safely before receiving feedback. Even though those drivers used the app less intensely, the scoring model is such that the marginal score increase of driving safer is larger for those with a lower score.

In addition to the data generated by the app, two questionnaires, at the beginning and at the end of our study, enabled us to provide evidence about potential selection and attrition biases. We thereby empirically confirmed concerns raised in the related literature (e.g. Elvik, 2014). More specifically, we found that novice drivers who drove daily and who had received traffic fines in the past were more likely to participate. Socio-economic characteristics did not matter much for this selection, possibly because the sample was fairly homogenous in terms of its (young) age.

Over the course of the field experiment, a substantial fraction of drivers decided to drop out by actively de-installing the app or by not using the app anymore. Reasons for dropping out seem to relate mostly to the functionality of the app (power usage, user-friendliness, and measurement precision of the underlying scoring model). Additionally, unlike active participants, dropouts exhibited on average little or no improvement in scores, even if measured on a trip basis. Since observed app use was similar between active participants and dropouts, this suggests that small improvements in scores may have induced participants to quit the field experiment prematurely. For future implementations of apps that intend to promote safe driving, these findings imply that app quality and usability are vital, and that a positive framing of (even small) improvements in driving skills, for instance by means of congratulatory messages, may be helpful in keeping users engaged.

Policymakers might be interested in implementing app-based feedback for novice drivers to improve safety and reduce the number of car crashes. Our study provides evidence for such an effect (through learning and monetary incentives) and guidance on which group is expected to exhibit improvements in their safety-relevant driving skills when given app-based feedback (and incentives). The app may also be valuable for pay-as-you-drive insurance schemes (in fact, related versions of the app employed here are already used by multiple European insurance companies for this purpose).

We suggest follow-up research on the effectiveness of telematics-based feedback for improving safety-relevant driving skills of (novice) drivers in order to overcome some of the limitations of this study (as discussed in Section 4.2). A larger sample size, a comprehensive control group and a consistent post-measurement for all participants would help in identifying the impact of telematics-based feedback on safety-relevant driving skills more concisely than this study allowed for. This could be achieved by providing higher incentives to participants (both, regular participants and members of the control group), preferably in the form of a fixed payment scheme for the "feedback only" treatment and a performance-dependent payment scheme for the "feedback & incentive" treatment. Compared to a lottery (as employed in this study), such payment schemes would render the relation between performance and payout more transparent to participants, potentially reducing selection

biases and confounding effects related to risk preferences. Furthermore, at least some part of the incentive should be conditional on having completed the post-measurement period, in order to make sure that participants do not drop out before the end of the post-measurement. Self-selection effects should be studied more thoroughly than this study allowed for, with

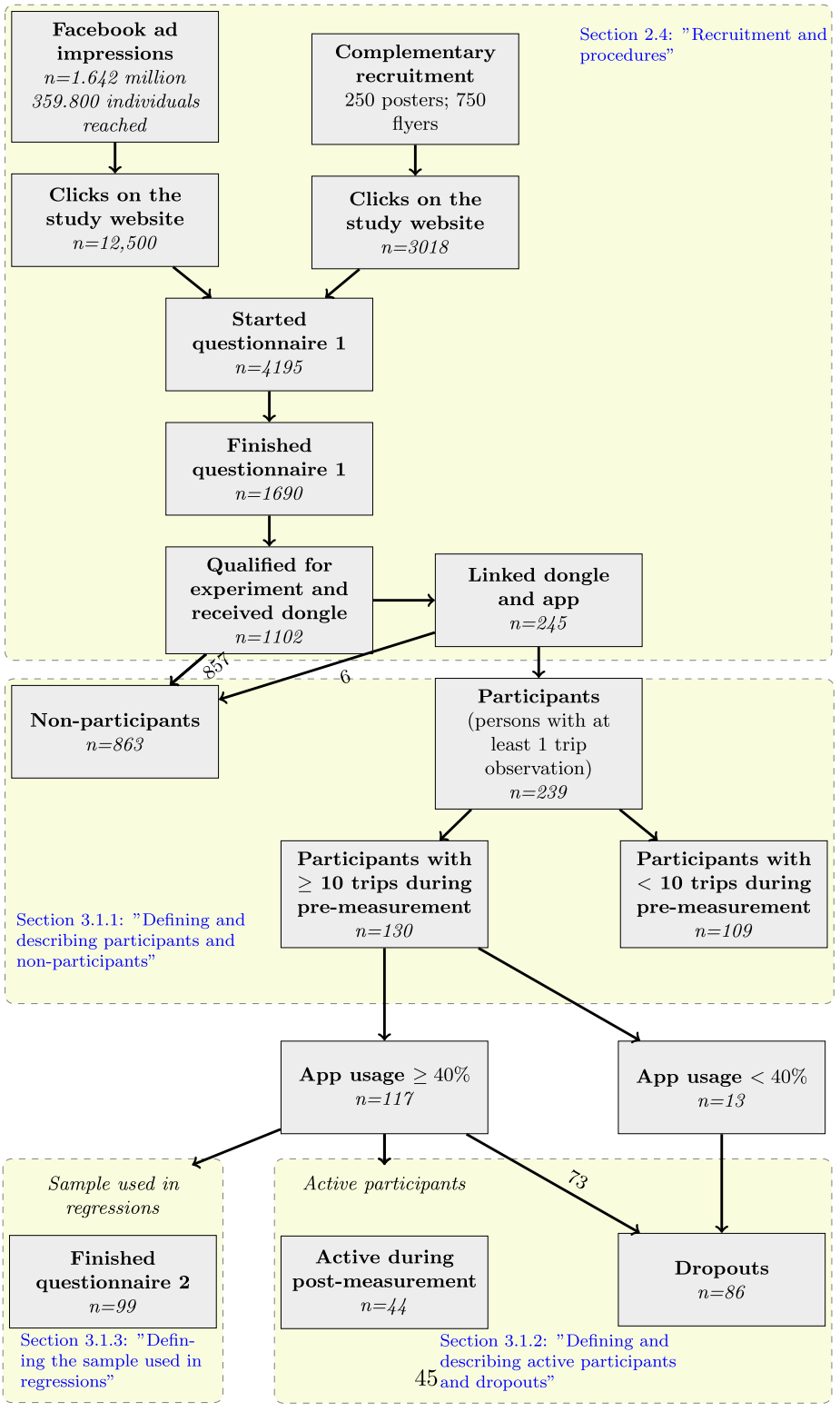


Fig. A.3. Diagram filtering down from the recruitment to the participants considered in the analyses.

the aim to shed light on why a large number of individuals targeted by advertisements did not react to them, and why a large share of those who visited the study website (∼15.500 persons) did not start the survey (∼11.300) or complete the survey (∼ 2.500). Reasons for this strong funnel effect in the recruitment should be identified and reduced in future studies. The same is true for factors that led to attrition in this study. These include the requirement to install both an on-board-diagnostic dongle and a smartphone app (and connect them to each other) as well as some software-related issues, such as the phone use indicator being defined differently for iOS and Android phones, and the high power usage of the smartphone app. With an improved incentive scheme and fewer factors inducing attrition, both self-selection and attrition biases are expected to decrease, providing a better indication of how the overall population of novice drivers would react to telematics-based feedback on their safety-relevant driving skills.

CRediT authorship contribution statement

Stefanie Peer: Conceptualization, Methodology, Writing – original draft. **Alexander Muermann**: Conceptualization, Methodology, Writing – review & editing. **Katharina Sallinger**: Data curation, Methodology, Formal analysis, Visualization, Writing – review & editing.

Acknowledgements

We wish to thank the anonymous reviewers for their constructive suggestions. We are also grateful to the participants of the hEART conference 2017, the ITEA conference 2018, and seminar participants at TU Dresden for valuable feedback, and the project consortium for their support of the study.

Appendix A. Overview sample

Fig. A.3

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.trf.2020.04.005.

References

Allouch, A., Koubâa, A., Abbes, T., & Ammar, A. (2017). Roadsense: Smartphone application to estimate road conditions using accelerometer and gyroscope. *IEEE Sensors Journal, 17*, 4231–4238.

Beloufa, S., Cauchard, F., Vedrenne, J., Vailleau, B., Kemeny, A., Mérienne, F., & Boucheix, J. M. (2017). Learning eco-driving behaviour in a driving simulator: contribution of instructional videos and interactive guidance system. *Transportation Research Part F: Traffic Psychology and Behaviour*..

Beusen, B., Broekx, S., Denys, T., Beckx, C., Degraeuwe, B., Gijsbers, M., ... Panis, L. I. (2009). Using on-board logging devices to study the longer-term impact of an eco-driving course. *Transportation Research Part D: Transport and Environment, 14*, 514–520.

Bolderdijk, J. W., Knockaert, J., Steg, E., & Verhoef, E. T. (2011). Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: Results of a Dutch field experiment. *Accident Analysis & Prevention, 43*, 1181–1186.

Botzer, A., Musicant, O., & Perry, A. (2017). Driver behavior with a smartphone collision warning application–a field study. *Safety Science, 91*, 361–372.

Braitman, K. A., Kirley, B. B., McCartt, A. T., & Chaudhary, N. K. (2008). Crashes of novice teenage drivers: Characteristics and contributing factors. *Journal of Safety Research, 39*, 47–54.

Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014). A meta-analysis of the effects of texting on driving. *Accident Analysis & Prevention, 71*, 311–318.

Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization, 131*, 141–150.

Chorlton, K., Hess, S., Jamson, S., & Wardman, M. (2012). Deal or no deal: Can incentives encourage widespread adoption of intelligent speed adaptation devices?. *Accident Analysis & Prevention, 48*, 73–82.

Chuang, M. C., Bala, R., Bernal, E. A., Paul, P., & Burry, A. (2014). Estimating gaze direction of vehicle drivers using a smartphone camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 165–170).

Cooper, P. J. (1997). The relationship between speeding behaviour (as measured by violation convictions) and crash involvement. *Journal of Safety Research, 28*, 83–95.

Curry, A. E., Hafetz, J., Kallan, M. J., Winston, F. K., & Durbin, D. R. (2011). Prevalence of teen driver errors leading to serious motor vehicle crashes. *Accident Analysis & Prevention, 43*, 1285–1290.

Dahl, R. E. (2008). Biological, developmental, and neurobehavioral factors relevant to adolescent driving risks. *American Journal of Preventive Medicine, 35*, S278–S284.

De Barba, P., Kennedy, G. E., & Ainley, M. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning, 32*, 218–231.

Dijksterhuis, C., Lewis-Evans, B., Jelijs, B., de Waard, D., Brookhuis, K., & Tucha, O. (2015). The impact of immediate or delayed feedback on driving behaviour in a simulated pay-as-you-drive system. *Accident Analysis & Prevention, 75*, 93–104.

Elvik, R. (2014). Rewarding safe and environmentally sustainable driving: systematic review of trials. *Transportation Research Record: Journal of the Transportation Research Board*, 1–7.

Etzioni, S., Erev, I., Ishaq, R., Elias, W., & Shiftan, Y. (2017). Self-monitoring of driving speed. *Accident Analysis & Prevention, 106*, 76–81.

Farah, H., Musicant, O., Shimshoni, Y., Toledo, T., Grimberg, E., Omer, H., & Lotan, T. (2013). The first year of driving: Can an in-vehicle data recorder and parental involvement make it safer? *Transportation Research Record, 2327*, 26–33.

Farah, H., Musicant, O., Shimshoni, Y., Toledo, T., Grimberg, E., Omer, H., & Lotan, T. (2014). Can providing feedback on driving behavior and training on parental vigilant care affect male teen drivers and their parents?. *Accident Analysis & Prevention, 69*, 62–70. https://doi.org/10.1016/j.aap.2013.11.005.

Farmer, C. M., Kirley, B. B., & McCartt, A. T. (2010). Effects of in-vehicle monitoring on the driving behavior of teenagers. *Journal of Safety Research, 41*, 39–45.

Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., & González, M. C. (2012). Safe driving using mobile phones. *IEEE Transactions on Intelligent Transportation Systems, 13*, 1462–1468.
Geyer, A., Kremslehner, D., & Muermann, A. (2019). Asymmetric information in automobile insurance: Evidence from driving behavior. *Journal of Risk and Insurance*..
Greaves, S., & Fifer, S. (2010). Development of a kilometer-based rewards system to encourage safer driving practices. *Transportation Research Record: Journal of the Transportation Research Board*, 88–96.
Grüninger, M. (2008). Das Unfallrisiko junger Fahrerinnen und Fahrer im geographischen Kontext: eine Auswertung der Unfallstatistik 2004 in Bayern. *Ifes*. https://www.ifes.fau.de/files/2017/07/GRUENINGER_2008_IfeS-Materialienband_3-2008.pdf.
Guttman, N., & Gesser-Edelsburg, A. (2011). "The little squealer" or "the virtual guardian angel"? Young drivers' and their parents' perspective on using a driver monitoring technology and its implications for parent-young driver communication. *Journal of Safety Research, 42*, 51–59.
Guttman, N., & Lotan, T. (2011). Spying or steering? views of parents of young novice drivers on the use and ethics of driver-monitoring technologies. *Accident Analysis & Prevention, 43*, 412–420.
Harré, N., Field, J., & Kirkwood, B. (1996). Gender differences and areas of common concern in the driving behaviors and attitudes of adolescents. *Journal of Safety Research, 27*, 163–173.
Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*, 1644–1655.
Holte, H. (2012). *Einflussfaktoren auf das Fahrverhalten und Unfallrisiko junger Fahrerinnen und Fahrer*. prefix https://www.bast.de/BASt_2017/DE/Publikationen/Berichte/unterreihe-m/2013-2012/m229.html. bASt-Bericht M-229..
Holte, H., Klimmt, C., Baumann, E., & Geber, S. (2014). *Wirkungsvolle Risikokommunikation für junge Fahrerinnen und Fahrer*. prefix https://www.bast.de/DE/Publikationen/Berichte/unterreihe-m/2015-2014/m249.html. bASt-Bericht M-249..
Horrey, W. J., Lesch, M. F., Dainoff, M. J., Robertson, M. M., & Noy, Y. I. (2012). On-board safety monitoring systems for driving: Review, knowledge gaps, and framework. *Journal of Safety Research, 43*, 49–58.
Huang, Y. H., Roetting, M., McDevitt, J. R., Melton, D., & Smith, G. S. (2005). Feedback by technology: Attitudes and opinions of truck drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 8*, 277–297.
Hultkrantz, L., & Lindberg, G. (2011). Pay-as-you-speed an economic field experiment. *Journal of Transport Economics and Policy (JTEP), 45*, 415–436.
Hurst, P. M. (1980). Can anyone reward safe driving?. *Accident Analysis & Prevention, 12*, 217–220.
Islam, S., Buttlar, W. G., Aldunate, R. G., & Vavrik, W. R. (2014). Measurement of pavement roughness using Android-based smartphone application. *Transportation Research Record, 2457*, 30–38.
Johnson, D. A. & Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. In *2011 14th International IEEE conference on Intelligent Transportation Systems (ITSC)* (pp. 1609–1615). IEEE..
Kervick, A. (2016). An evaluation of smartphone driver support systems for young drivers-acceptance, efficacy, and driver distraction. Ph.D. thesis. NUI Galway. prefix https://aran.library.nuigalway.ie/handle/10379/6091..
Kervick, A. A., Hogan, M. J., O'Hora, D., & Sarma, K. M. (2015). Testing a structural model of young driver willingness to uptake smartphone driver support systems. *Accident Analysis & Prevention, 83*, 171–181.
Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine, 370*, 54–59.
Lahrmann, H., Agerholm, N., Tradisauskas, N., Næss, T., Juhl, J., & Harms, L. (2012). Pay as you speed, ISA with incentives for not speeding: A case of test driver recruitment. *Accident Analysis & Prevention, 48*, 10–16.
Laiou, A., Yannis, G., Folla, K., Bauer, R., Machata, K., & Brandstaetter, C. (2010). An overview of road accident fatalities in the. *European Union*. https://www.nrso.ntua.gr/geyannis/wp-content/uploads/geyannis-pc267.pdf.
Lehmann, G. & Cheale, A. (1998). The contribution of onboard recording systems to road safety and accident analysis. In *16th ESV conference, paper*..
Lerner, N., Jenness, J., Singer, J., Klauer, S., Lee, S., Donath, M., ... Ward, N. (2010). *An exploration of vehicle-based monitoring of novice teen drivers: Final report (DOT HS 811 333)*. Washington, DC: National Highway Traffic Safety Administration. https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/811333.pdf.
Lotan, T., Musicant, O., & Grimberg, E. (2014). *Can young drivers be motivated to use smartphone based driving feedback?–A pilot study*. Technical Report.
Mazureck, U., & Hattem, J. (2006). Rewards for safe driving behavior: Influence on following distance and speed. *Transportation Research Record: Journal of the Transportation Research Board*, 31–38.
McKnight, A. J., & McKnight, A. S. (2003). Young novice drivers: Careless or clueless? *Accident Analysis & Prevention, 35*, 921–925.
Meseguer, J. E., Calafate, C. T., Cano, J. C., & Manzoni, P. (2013). Drivingstyles: A smartphone application to assess driver behavior. In *2013 IEEE Symposium on Computers and Communications (ISCC)* (pp. 000535–000540). IEEE..
Mullen, N. W., Maxwell, H., & Bedard, M. (2015). Decreasing driver speeding with feedback and a token economy. *Transportation Research Part F: Traffic Psychology and Behaviour, 28*, 77–85.
Musicant, O., & Lotan, T. (2016). Can novice drivers be motivated to use a smartphone based app that monitors their behavior?. *Transportation Research Part F: Traffic Psychology and Behaviour, 42*, 544–557.
Neyens, D. M., & Boyle, L. N. (2007). The effect of distractions on the crash types of teenage drivers. *Accident Analysis & Prevention, 39*, 206–212.
Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., & Summala, H. (2006). Cross-cultural differences in driving behaviours: A comparison of six countries. *Transportation Research part F: Traffic Psychology and Behaviour, 9*, 227–242.
Paefgen, J., Kehr, F., Zhai, Y., & Michahelles, F. (2012). Driving behavior analysis with smartphones: Insights from a controlled field study. In *Proceedings of the 11th international conference on mobile and ubiquitous multimedia* (pp. 36).
Reagan, I. J., Bliss, J. P., Van Houten, R., & Hilton, B. W. (2013). The effects of external motivation and real-time automated feedback on speeding behavior in a naturalistic setting. *Human Factors, 55*, 218–230.
Roetting, M., Huang, Y. H., McDevitt, J. R., & Melton, D. (2003). When technology tells you how you drive—-truck drivers' attitudes towards feedback by technology. *Transportation Research Part F: Traffic Psychology and Behaviour, 6*, 275–287.
SafetyNet (2009). Novice drivers. Technical Report. https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/specialist/knowledge/pdf/novice_drivers.pdf..
Sarma, K. M., Carey, R. N., Kervick, A. A., & Bimpeh, Y. (2013). Psychological factors associated with indices of risky, reckless and cautious driving in a national sample of drivers in the Republic of Ireland. *Accident Analysis & Prevention, 50*, 1226–1235.
Shawky, M., Al-Badi, Y., Sahnoon, I., Al-Harthi, H. (2017). The relationship between traffic rule violations and accident involvement records of drivers. In *Advances in human aspects of transportation* (pp. 745–755). Springer..
Simons-Morton, B. G., Ouimet, M. C., Zhang, Z., Klauer, S. E., Lee, S. E., Wang, J., ... Dingus, T. A. (2011). The effect of passengers and risk-taking friends on risky driving and crashes/near crashes among novice teenagers. *Journal of Adolescent Health, 49*, 587–593.
Soriguera, F., & Miralles, E. (2016). Driver feedback mobile app. *Transportation Research Procedia, 18*, 264–271.
Statistik Austria (2017). Strassenverkehrsunfälle mit Personenschaden, Jahresergebnisse 2016. https://www.statistik.at/web_de/services/publikationen/14/index.html?includePage=detailedView&sectionName=Verkehr&pubId=751..
Statistik Austria (2018). Straßenverkehrsunfälle mit Personenschaden. Jahresergebnisse 2017. https://www.statistik.at/web_de/services/publikationen/14/index.html?includePage=detailedView&sectionName=Verkehr&pubId=766..
Taubman-Ben-Ari, O., & Yehiel, D. (2012). Driving styles and their associations with personality and motivation. *Accident Analysis & Prevention, 45*, 416–422.
Toledo, T., Lotan, T., Taubman-Ben-Ari, O., & Grimberg, E. (2012). Evaluation of a program to enhance young drivers' safety in Israel. *Accident Analysis & Prevention, 45*, 705–710.

Toledo, T., Musicant, O., & Lotan, T. (2008). In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transportation Research Part C: Emerging Technologies, 16*, 320–331.
Tulusan, J., Staake, T., & Fleisch, E. (2012). Providing eco-driving feedback to corporate car drivers: what impact does a smartphone application have on their fuel efficiency?. In *Proceedings of the 2012 ACM conference on ubiquitous computing ACM* (pp. 212–215).
Turner, C., & McClure, R. (2003). Age and gender differences in risk-taking behaviour as an explanation for high incidence of motor vehicle crashes as a driver in young males. *Injury Control and Safety Promotion, 10*, 123–130.
Ulleberg, P., & Rundmo, T. (2003). Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Safety Science, 41*, 427–443.
Vaiana, R., Iuele, T., Astarita, V., Caruso, M. V., Tassitani, A., Zaffino, C., & Giofrè, V. P. (2014). Driving behavior and traffic safety: An acceleration-based safety evaluation procedure for smartphones. *Modern Applied Science, 8*, 88.
Vegega, M., Jones, B., & Monk, C. (2013). *Understanding the effects of distracted driving and developing strategies to reduce resulting deaths and injuries: a report to Congress* (Report No. DOT HS 812 053). Washington, DC: National Highway Traffic Safety Administration.
Wåhlberg, A. E. (2007). Long-term effects of training in economical driving: Fuel consumption, accidents, driver acceleration behavior and technical feedback. *International Journal of Industrial Ergonomics, 37*, 333–343.
Williams, A. F. (2003). Teenage drivers: Patterns of risk. *Journal of Safety Research, 34*, 5–15.
Wouters, P. I., & Bos, J. M. (2000). Traffic accident reduction by monitoring driver behaviour with in-car data recorders. *Accident Analysis & Prevention, 32*, 643–650.
Young, K. L., Regan, M. A., Triggs, T. J., Jontof-Hutter, K., & Newstead, S. (2010). Intelligent speed adaptation—effects and acceptance by young inexperienced drivers. *Accident Analysis & Prevention, 42*, 935–943.