Peer, Stefanie; Knockaert, Jasper; Verhoef, Erik

Working Paper

Train Commuters' Scheduling Preferences: Evidence from a Large-Scale Peak Avoidance Experiment

Tinbergen Institute Discussion Paper, No. 15-078/VIII

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Peer, Stefanie; Knockaert, Jasper; Verhoef, Erik (2015) : Train Commuters' Scheduling Preferences: Evidence from a Large-Scale Peak Avoidance Experiment, Tinbergen Institute Discussion Paper, No. 15-078/VIII, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/125079

TI 2015-078/VIII
Tinbergen Institute Discussion Paper

Train Commuters' Scheduling Preferences: Evidence from a Large-Scale Peak Avoidance Experiment

Stefanie Peer[a,b,c]
Jasper Knockaert[a]
Erik Verhoef[a,b]

[a] *Faculty of Economics and Business Administration, VU University Amsterdam, the Netherlands;*
[b] *Tinbergen Institute, the Netherlands;*
[c] *Vienna University of Economics and Business, Vienna, Austria.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Train commuters' scheduling preferences: evidence from a large-scale peak avoidance experiment[☆]

Stefanie Peer[a,c,*], Jasper Knockaert[a], Erik T. Verhoef[a,b]

[a] *Department of Spatial Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam*
[b] *Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam*
[c] *Department of Socioeconomics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna*

Abstract

We study the trip scheduling preferences of train commuters in a real-life setting. The underlying data have been collected during large-scale peak avoidance experiment conducted in the Netherlands, in which participants could earn monetary rewards for traveling outside peak hours. The experiment included ca. 1000 participants and lasted for about 6 months. Holders of an annual train pass were invited to join the experiment, and a customized smartphone app was used to measure the travel behavior of the participants. We find that compared to the pre-measurement, the relative share of peak trips decreased by 22% during the reward period, and by 10% during the post-measurement. By combining multiple complementary data sources, we are able to specify and estimate (MNL and panel latent class) departure time choice models. These yield plausible estimates for the monetary values that participants attach to reducing travel time, schedule delays, the number of transfers, crowdedness, and unreliability.

Keywords: departure time decisions, scheduling, peak avoidance experiment, rail, trains, revealed preference data, smartphone app, GPS data, value of travel time, valuation of schedule delays, valuation of comfort, crowding costs, valuation of reliability, costs of transfers
JEL codes: C25, C90, D01, D80, R41

[☆]The authors would like to thank the participants of the 2013 conference of the International Transportation Economic Association (ITEA), the 2014 conference of the European Regional Science Association (ERSA), the 2014 conference of the European Association for Research in Transportation (hEART) and seminar participants at the KTH University in Stockholm and the Masaryk University in Brno for useful comments and suggestions. The Dutch Ministry of Infrastructure and the Environment is thanked for their financial support of this study. Also financial support from the ERC (Advanced Grant OPTION #246969) for the research of Erik Verhoef and from the NWO (SAR project iPriSM) for the research of Jasper Knockaert is gratefully acknowledged. The usual disclaimer applies.

[*]Corresponding author: `speer@wu.ac.at` (Stefanie Peer)

1. Introduction

The Netherlands, being one of the most densely populated countries, has one of the busiest railway networks.[1] Recent trends suggest that demand for train travel will still grow further: between 2000 and 2012, the number of railway passenger kilometers rose by 19%, while during the same period car travel increased only by 3% (Kennisinstituut voor Mobiliteitsbeleid (KiM), 2013, pp.32,39). The modal share of trains is particularly high for medium- and long-distance commutes and for commutes between the major Dutch cities (Schwanen et al., 2002). Especially in the Western and central parts of the Netherlands (comprising the cities of Amsterdam, Rotterdam, The Hague, Leiden and Utrecht) the supply of train services is close to capacity during peak hours, both in terms of network and train capacity (van Vuuren, 2002). The resulting crowdedness leads to discomfort among passengers, mainly due to a lack of empty seats and personal space (e.g. Li and Hensher, 2011). Moreover, the large number of people boarding and leaving the trains may lead to longer travel times as well as lower levels of reliability.

Clearly, crowdedness can be reduced by increasing (train and rail network) capacity. However, capacity expansion may not be the most cost-efficient solution, especially when demand is strongly concentrated during peak hours, meaning that the additional capacity will be largely idle outside the peak periods.[2] An alternative to expensive additions to the rolling stock and infrastructure amendments are fares that differ by time of day. The basic idea is that discounted off-peak tickets induce some travelers to shift their trips from peak to off-peak periods, hence decreasing crowdedness during peak hours – while avoiding the costs that would accrue for capacity expansions.

Such time-differentiated charging schemes tend to be beneficial not only from the perspective of the train operators, but also from a social welfare perspective. The main reason is that higher fares during the peak periods lead to an internalization of the external crowding costs, which the passengers impose on each other. In the context of car travel, time-differentiated pricing schemes have been investigated and advocated by transport economists since the publication of the seminal paper by Vickrey (1969). In his paper, Vickrey introduced the well-known bottleneck model, which assumes that travelers make a trade-off between travel time and schedule delays (i.e. the disutility of arriving at the destination earlier or later than preferred) when deciding on their departure time. If the preferred arrival times are similar across travelers (for instance due to common working hours), demand may exceed road capacity and as a consequence congestion will form. In equilibrium, those drivers who arrive earlier or later than preferred will have shorter travel times and longer schedule delays than those that arrive close to their preferred arrival time.

In Vickrey's modeling framework, under first-best conditions, congestion can be fully eliminated by using tolls that vary by time of day. A recent paper by De Palma et al. (2013) shows that similar results can be derived for public transport when a trade-off between schedule delays and crowding applies. The application of an optimal charging scheme will then lead to a more uniform distribution of travelers between train connections compared to the user equilibrium, however, without fully eliminating crowding. Moreover, time-differentiated fares might also be beneficial from an environmental perspective: Rietveld (2002) argues that the marginal environmental costs are higher for peak train travelers than for off-peak train travelers when the off-peak supply of

[1] See for instance here: `http://www.cbs.nl/en-GB/menu/themas/verkeer-vervoer/publicaties/artikelen/archief/2009/2009-2702-wm.htm`

[2] In the Netherlands, the concentration of train trips during peak periods is rather high: 24% of all public transport trips longer than 10 km (a majority of which are made by train) take place during the morning peak, but only 13% of all car trips (Kennisinstituut voor Mobiliteitsbeleid (KiM), 2013, p.38).

train capacity is determined to a large extent by the demand during the peak period. Time-differentiated fares that are higher during peak periods will then contribute to the internalization of the environmental costs.

The Dutch National Railways (NS)[3] recognized already in the late 1980ies that time-of-day-dependent ticket prices may be a more cost-efficient alternative to expensive capacity expansions. They introduced the so-called 'Rail-Aktief-Kaart', which granted a 40% discount on single tickets for off-peak trips and still exists today (under the name 'Dal Voordeel abonnement'). While the 'Rail-Aktief-Kaart' allows for time-differentiated pricing for single trips, the 'peak avoidance experiment' ('Spitsmijden' in Dutch) discussed in this paper is targeted at regular Dutch train travelers (typically commuters) who hold annual train passes. At the core of the experiment is a (distance-dependent) reward scheme, which grants monetary rewards for each trip outside the morning and evening peak hours. The peak avoidance experiment took place between summer 2012 and spring 2013, including more than 1000 active participants. Pre-condition for participation was a valid annual travel pass for a specific OD-pair or the entire Dutch railway network. The travel behavior of the participants was measured via a customized smartphone app, which continuously recorded the global positioning system (GPS) coordinates, and hence allowed for a direct measurement of their travel behavior.

Besides the reward period of 22–25 weeks, the travel behavior of the participants was also measured during a 3-week period of pre-measurement and a 4-week period of post-measurement. This renders it possible to identify the effect of the reward incentive on the travel behavior of the participants. In order to get the rewards paid out, the participants additionally had to fill in various surveys as well as logbooks. These complementary, high-quality data sources are used to gain further insights into the socio-economic characteristics of the participants, and their (usual and preferred) scheduling behavior and preferences. Together with the travel information provided by an app that is made available by the Dutch National Railways (NS), they render it feasible to estimate full-fledged scheduling choice models, including definitions of the following attributes of the departure time choice alternatives: rewards, travel time, schedule delays, number of transfers, crowdedness (as a proxy for comfort) and unreliability. Due to the presence of the time-of-day-dependent rewards, the monetary valuations associated with improvements in these attributes can be derived. Both multinomial logit and panel latent class models are estimated.

So far, most existing research on the monetary valuation of train travel attributes[4] has focused only on a small number of travel attributes, resulting in models that are less comprehensive than the ones presented in this paper. For instance, Douglas and Karpouzis (2006), Basu and Hunt (2012) and the meta-analysis of Wardman and Whelan (2011) concentrate on crowding in trains, Liu et al. (1997) on transfers and Rietveld et al. (2001) on travel time reliability. A more comprehensive picture on valuations in train travel can be obtained from the meta-analyses on public transport valuations by Wardman (2001, 2004). These two papers by Wardman provide an overview of a large number of (mainly British) public transport valuation studies, accounting for most of the relevant travel attributes (except for comfort valuations). To a large degree these valuations are reported separately for train travel (as opposed to other modes), rendering them especially relevant in the context of this paper.

[3] Nederlandse Spoorwegen (`www.ns.nl`)

[4] Extensive literature exists on the valuation of travel attributes in general, especially in the context of road transport. Recent meta-analyses can for instance be found in Shires and De Jong (2009) (on the value of time), Carrion and Levinson (2012) (on the valuation of reliability) and Li and Hensher (2011) (on the valuation of comfort).

Most studies that enter the meta-analyses of Wardman as well as all the other above-cited papers use stated preference (SP) rather than revealed preference (RP) data.[5] Unlike the models presented in this paper, those valuations are thus derived from the behavior of travelers in hypothetical rather than real-life choice situations, which may lead to a significant divergence in the resulting valuations, as found by Wardman (2001, 2004). Among the few existing RP studies on valuations of train travel attributes are papers focusing on reliability by van Loon et al. (2011), on transfers by Guo and Wilson (2011), as well as a mode choice study based on combined SP-RP data by Polydoropoulou and Ben-Akiva (2001). However, these studies either use aggregate data rather than data from individual travelers (e.g. van Loon et al., 2011), or individual data without panel structure (e.g. Polydoropoulou and Ben-Akiva, 2001; Guo and Wilson, 2011). In contrast, our analysis is based on panel data of individual scheduling decisions, and thus allows drawing conclusions on how individual travelers change their behavior over time, in particular as a response to the introduction of rewards for off-peak travel. Given that this study is one of few to estimate scheduling choice models using RP data in a public transport context, we discuss specific issues that arise under these circumstances, and propose methods to deal with them. These issues include among others the specification of the choice set, the definition of attribute values for non-chosen alternatives and the selection of relevant observations.

We also add to that branch of the transport economics literature that focuses on changes in travel patterns due the introduction of time-differentiated tolls, fares or rewards, including studies on the effects of the congestion charging schemes in London (e.g. Santos and Shaffer, 2004) and Stockholm (e.g. Eliasson et al., 2009; Karlström and Franklin, 2009). Knockaert et al. (2011) provide an overview of earlier peak avoidance ('Spitsmijden') experiments conducted in the Netherlands, in which monetary rewards for off-peak travel were granted. While most of these experiments were targeted at car travelers, there is one exception: a small-scale experiment with the aim to test how train pass holders react to time-differentiated pricing (Samenwerkingsverband Spitsmijden, 2009). Although the focus of that experiment was similar to the one described in this paper, it was considerably smaller in scale (involving only 124 participants), and the logbook entries of the participants were used to measure travel behavior (rather than actual travel time measurements from GPS data or similar sources). It turned out that already before the start of the experiment, the majority of the participants traveled mostly outside the peak, and during the experiment all trips were reported to take place outside the peak. The results were thus only to a very limited degree generalizable. Yet another study on time-of-day-dependent fares in train travel has been conducted by Currie (2010). Using aggregate data on travel behavior, he describes the effect of providing free "early-bird tickets" to rail passengers in Melbourne who complete their travel by 7:00 a.m. He concludes that this new pricing policy reduces demand during the peak periods significantly and suggests that the provision of the free tickets is budget neutral (as fewer trains are needed during peak periods).

Finally, the paper also adds to the fairly recent literature on the possibilities to use GPS data for measuring travel behavior (e.g. Schönfelder et al., 2002; Duncan et al., 2009; Houston et al., 2014). While the existing literature contains GPS applications concerning car travel (e.g. Murakami and Wagner, 1999; Peer et al., 2013), walking (e.g. Kang et al., 2013), cycling (e.g. Broach et al., 2012) and bus travel (e.g. Lin and Zeng, 1999; Mazloumi et al., 2010), we are not aware of any

[5]The predominance of SP data is generally true for the literature on the valuation of travel attributes. Among those studies that are (at least partially) based on RP data are Lam and Small (2001), Brownstone and Small (2005) and Börjesson (2008).

application customized to train travel, except for various studies that focus on the identification of travel modes from GPS data (e.g. Bohte and Maat, 2009; Gong et al., 2012). The lack of studies covering GPS applications in the train travel domain may be related to the specific difficulties that arise there, most of which are related to the poor reception of satellite signals in trains and stations (Bohte and Maat, 2009; Stopher et al., 2005). This paper suggests data processing procedures that tackle these difficulties.

The paper unfolds as follows. Section 2 discusses the design of the peak avoidance experiment and the recruitment of the participants. In Section 3, the decision to participate in the experiment as well as the decision to drop out of the experiment are investigated. Section 4 gives an overview of the collected data and discusses how these data can be used to measure the travel behavior of the participants. It proceeds with providing descriptive statistics on the changes in travel behavior induced by the reward for off-peak trips. Section 5 introduces the behavioral models that are used to derive the monetary valuations for the travel attributes and provides the corresponding estimation results. Finally, Section 6 concludes.

2. Experimental Design

2.1. General setup

The peak avoidance experiment covered in this paper took place in the Netherlands between August 2012 and April 2013, with the aim to quantify how the scheduling behavior of regular train users changes when off-peak travel is rewarded. The travel behavior of the participants was measured using the GPS data recorded by a customized smartphone app, which the participants had to use and which is described in more detail in Section 4.1.

Owners of an annual train pass for specific railway routes or for the entire Dutch railway network were eligible to participate in the experiment (the recruitment strategy will be discussed in more detail in Section 2.2). Participants could initiate their participation at three different dates. Depending on their starting date, their participation in the experiment was planned to last between 22 to 25 weeks. The first 3 weeks of participation constituted the pre-measurement, the subsequent 15-18 weeks the reward period and the final 4 weeks constituted the post-measurement. During the pre- and post-measurement, off-peak travel was unrewarded. Thanks to these reference periods, we are able to quantify to which extent the scheduling behavior of the participants changes when off-peak travel gets rewarded.

All participants had to indicate a so-called 'participation OD-pair' (so the origin-destination pair for which they wanted to participate to the experiment) when they registered for the experiment. For those who had a route-specific pass, the participation OD-pair was generally identical to the route for which they had the pass. Those with passes for the entire network were able to choose a participation OD-pair. During the reward period, the participants were able to gain a reward if they started their trip from the begin station to the end station of their participation OD-pair outside the morning peak (6:30-9:00), and if they started their trip from the end station to the begin station of their participation OD-pair outside the evening peak (16:00-18:30). The participants could thus receive a maximum of two rewards per day. Rewards were only distributed during weekdays, not during weekends.

Note that the participants received a *reward for making an app-registered trip along their participation OD-pair outside the peak hours*. This implies that they were not rewarded for working from home or for choosing other modes of transport. This setup was chosen to prevent that

participants have an incentive to keep their train trips unregistered, for instance by leaving their phone at home or switching it off. For similar reasons, the number of rewards that can be earned are independent from the observed travel behavior during the pre-measurement.[6] If that was not the case, participants would have an incentive to not have their off-peak trips registered during the pre-measurement.

The reward level is dependent on the length of the participation OD-pair, as shown in Table 1. Moreover, each participant is randomly assigned a low or a high reward level for the first half of the reward period, and gets assigned the other level for the second half of the reward period. This renders it possible to investigate the influence of the reward level on the scheduling behavior of the participants.

In order to actually receive the rewards, the participants had to fill in three surveys and six logbooks:

1. **Initial survey**
   In this first survey, participants were asked to provide information on their socio-economic characteristics, their scheduling restrictions and their preferred arrival and departure times. Moreover, they were asked to report their usual travel behavior (and the characteristics thereof).

2. **Intermediary survey** In the second survey a hypothetical scheduling choice experiment was included. Participants were also asked to provide information on their employer and their motivation for participation.

3. **Final survey** The last survey contained evaluation questions regarding various aspects of the experiment as well as questions about the expected future scheduling behavior. Furthermore, participants had to fill in a hypothetical ticket choice experiment.

4. **Logbooks** The participants had to fill in 6 logbooks, each of them covering a period of 1 week: 1 during the pre-measurement, 4 during the reward period and 1 during the post-measurement. For each weekday during those 6 weeks, they had to provide detailed information regarding their travel behavior. They were asked to indicate whether, when and where they traveled on a given day and which travel mode they chose. If they traveled by train,

[6] For instance, a person who traveled off-peak 3 times per week both during the pre-measurement and the reward period will receive the same number of rewards as a person who traveled off-peak 0 times per week during the pre-measurement and 3 times during the reward period.

Table 1: Reward scheme

	Reward per off-peak trip	
Travel distance	high	low
$\leq$ 25 km	2.5 Euro	1.5 Euro
26–40 km	3.5 Euro	2.5 Euro
$>$ 40 km	4.5 Euro	2.5 Euro

they moreover had to indicate if (and by how much) the train was delayed and how crowded it was. Finally, the participants were asked to grade their overall train travel experience for each of the 6 weeks.

The logbooks do not only provide additional data on the travel behavior of individual participants, which cannot be collected via the smartphone app, but they also allow for a continuous monitoring of many participation OD-pairs: since the setup of the logbook schedule was such that for any week between summer 2012 and spring 2013 one group of participants was asked to fill in the logbook, we have continuous data for those OD-pairs that are used by a sufficiently high number of participants.

The rewards for off-peak trips were transferred to the participants in two parts: in the middle of the reward period and at the end of the post-measurement. To receive the rewards, the participants had to have filled in all surveys and logbooks that were made available to them until each of the two payment dates.

2.2. Recruitment

The participants were recruited from existing clients of all Dutch (passenger) railway companies during two recruitment rounds (the first one taking place in the summer, and the second one in late autumn 2012). Only if they already owned an annual rail pass for a specific OD-pair or the entire Dutch network, they were entitled to participate. This restriction follows from the experiment's focus on travelers who use the train to commute to work (or, in some cases, to educational institutions). As a result, 86065 eligible persons[7] received a personal invitation via email during one of the two recruitment rounds. During the second recruitment round, in addition to the emails, also posters at the most frequented stations of the Netherlands were used to approach potential participants.

The two recruitment rounds resulted in 1009 participants who downloaded and used the app successfully.[8] This implies a response rate of 1.2%, if only those approached via email are considered. And it is even lower if one takes into consideration that some train travelers may have registered as a consequence of having seen the posters at the stations rather than having received an email. Clearly, the response rate is rather low, also compared to other peak avoidance experiments, where response rates between 5-20% (Samenwerkingsverband Spitsmijden, 2009; Knockaert et al., 2012a,b) were achieved.

One possible explanation for the low response rate is the recruitment method. In contrast to most previous peak avoidance experiments, potential participants were approached by email rather than by postal mail. A large number of recipients may have ignored the emails, expecting that they contain advertisements for train journeys.[9] Other reasons for declining the invitation can be derived from the outcomes of the final survey. In this survey, 38% of those who ultimately decided

[7] A large number of pass holders could not be approached because they had indicated that they do not want to receive commercial mailings or requests for (market) research from the railway companies. Moreover, in cases where the pass was (partially) paid by the employer, the employers could forbid the railway companies to contact their employees directly.

[8] Indicated by at least one valid registration of a train commute.

[9] The recruitment of respondents for a questionnaire targeted at non-participants (see Section 3.1) points in the same direction. Non-participants received an email that offered them a generous reward of 20 Euro if they filled in a survey that would last no longer than 20 minutes. The response rate was 13% in this case, which is still rather low considering the financial attractiveness of the offer.

to participate stated that they had doubts regarding their participation. One main cause for their doubts was the technical complexity of the experiment, including the requirement to have access to a suitable smartphone (iOS or Android platform, preferably with unlimited internet access), to register online and to download the app. Some participants also reported that their doubts were raised as they were unsure about the setup of the experiments and the rewards they could gain. Other reported sources of doubt included the time requirements to fill in the surveys and logbooks as well as privacy concerns (e.g. whether the GPS signals collected via the smartphone app would stay confidential).

3. Analysis of participation behavior

3.1. Participants vs. non-participants

The participants of the experiment are likely to constitute a non-representative sample of Dutch train commuters. Specifically, we expect that participants were able to gain rewards with less effort than non-participants (i.e. those who were invited to participate but decided against participation).

In this section, we will thus compare participants and non-participants in terms of their socio-economic characteristics, scheduling restrictions and usual travel patterns. This comparison does not only allow us to obtain information on the representativeness of the participants with respect to the non-participants, but also on the determinants of the participation decision.

For participants, the relevant data were collected in the initial and intermediary survey. Overall, 649 participants filled in both surveys. Non-participants were approached via email with a questionnaire that contained a selection of the participants' survey questions as well as some extra questions. Among the 3697 recipients of the email, 489 filled in the questionnaire (implying a response rate of ca. 13%) and received a voucher worth 20 Euro.

Table 2 shows the results of the comparison between participants and non-participants along various dimensions. The two groups show to be rather similar in terms of age, number of people living in the same household as well as income. Larger differences exist with respect to the education level: participants are on average more highly educated than non-participants. For instance, among participants 42.5% have a university degree, vs. 25.4% among non-participants. It is also evident from Table 2 that an over-proportionally large number of participants work as researchers, engineers, teachers and specialists, whereas they are underrepresented in administrative, service and sales jobs. This already gives an indication that participants tend to have jobs that are usually associated with fairly flexible working hours, making it easier for them to avoid the peak. And that indication is confirmed in the answers of participants and non-participants to questions that are directly related to the scheduling restrictions they face. Participants state to have more flexibility in terms of working (start/end) hours. Restrictions imposed by their employers as well as by their personal situation are less stringent.

Table 2: Comparing participants and non-participants

	Participants	Non-participants
Number	649	489
Average age	39.7	41.7
Average number of persons per household	2.7	2.7

Monthly household net income		
< 2000 Euro	10.3%	8.6%
$< 2000 - 3500$ Euro	29.6%	32.3%
> 3500 Euro	32.2%	29.0%
Unknown	27.9%	30.1%
Education (highest degree)		
Secondary education degree or lower	21.4%	31.1%
College / applied university degree	35.0%	41.9%
University degree	42.5%	25.4%
Other	1.1%	1.6%
Job type		
Manager	12.6%	10.2%
Researcher, engineer, teacher, specialist	54.4%	35.2%
Administrative, service and sales personnel	30.1%	54.0%
Other	2.9%	0.6%
Scheduling restrictions imposed by employer		
Can start work immediately upon (early) arrival	85.2%	83.2%
Can start work later than usual	82.1%	71.6%
Can end work later than usual	91.8%	87.7%
Scheduling restrictions imposed by employee		
Can leave home earlier than usual	80.3%	69.9%
Can leave home later than usual	89.4%	71.0%
Can leave work earlier than usual	82.3%	61.9%
Can leave work later than usual	87.8%	72.0%
Usual departure times (before experiment)		
Morning: before 6:30	19.1%	15.1%
Morning: after 9:00	6.6%	1.2%
Evening: before 16:00	12.6%	5.3%
Evening: after 18:30	8.9%	1.8%

Finally, differences also exist in the (reported) usual timing of the commute trips before participation in the experiment. A substantially higher share of participants than non-participants indicates to usually travel outside the peak hours also when no reward for off-peak travel is granted. While not reported, the same is true for the time periods at the edges of the peak periods, in which participants tend to be overrepresented as well. And furthermore, also the preferred arrival times of the participants tend to be closer to the off-peak periods than for non-participants.

The empirical evidence thus supports our expectation that participants can on average earn the rewards with less effort compared to non-participants, mainly because participants tend to be more flexible and they tend to travel close to or during off-peak periods also in absence of a monetary incentive for off-peak travel.

3.2. Participants vs. dropouts

A significant share of the 1009 initial participants quitted before the end of their participation period of 22-25 weeks. As Table 3 shows, 467 (46.3%) out of the initial 1009 participants filled in all obligatory surveys and logbooks, and 478 (47.4%) had the customized app (which is required for trip registration) still switched on during the post-measurement. Not filling in the obligatory surveys and logbooks meant that the rewards would not be paid out, whereas turning off the app during the post-measurement had no immediate consequences for the participants.

One of the main reasons for participants to drop out (or at least for switching off the app during the post-measurement) was the experimental character of the customized app. No less than 40% of the participants evaluated the app as negative or rather negative at the end of the experiment. The app was meant to be user-friendly by running passively in the background, hence without the need to switch it on and off before and after train trips, respectively. Unfortunately, this setup induced a rather high power usage, which made it impossible for many participants to have the app turned on the entire day without re-charging the phone during the day. Moreover, also the trip registration did not always work as intended (for instance due to tunnels and low-quality internet connections): in the final survey only 19% of the participants state that all their trips had been registered correctly. As a result some of them forewent rewards, which they would have been entitled to receive. The remaining 81% of participants estimate that on average 71% of their trips were registered correctly (meaning that the begin and end station as well as the corresponding departure and arrival times were registered correctly).[10] Also other technical issues were reported, such as the app shutting down without further notification.[11]

Other explanations for the rather high number of drop-outs can be found by comparing those who filled in all six logbooks to those who filled in at least one, but not all logbooks. These two groups represent those who participated until the end of the experiment (filling in all logbooks was

[10]The participants had the possibility to contact a help desk and have their trips being registered manually if (i) the GPS data provided sufficient evidence of their asserted travel behavior or if (ii) they had other evidence supporting their claims. However, all analyses presented in this paper are based on the original data, rather than the manually corrected ones.

[11]The technical problems were seemingly worse for Android users. Only 54% of them judge the app as positive (or rather positive), compared to 67% of iOS users. Moreover, only 14% of Android users report that all their trips had been registered correctly, compared to 24% of iOS users. Finally, 84% of the Android users agree that the battery discharges quickly if the app is switched on, compared to 53% of iOS users.

Table 3: Overview recruitment

Description	Numbers
Personally invited for participation	86065
Participants with at least one app-registered train commute trip*	1009 (1.2% of 86065)
Participants who filled in all surveys and logbooks	467 (46.3% of 1009)
Participants with at least one app-registered train commute trip during post-measurement*	478 (47.4% of 1009)

*Train commute trips are defined as described in Section 4.2.

a requirement to receive the rewards) and those who participated actively at the beginning, but dropped out sometime during the experiment.

Table 4 shows that participants who decided to drop out prematurely belong relatively more often to the lowest distance class and relatively less often to the highest distance class than those who completed the experiment. This implies that those who quitted prematurely obtained on average a lower reward per off-peak trip, while facing the same duties in terms of completing the logbooks and surveys compared to those who participated until the end of the experiment.

Also the travel behavior between these two groups differs. For those who dropped out we observed a lower number of average weekly train commute trips than for those who completed the experiment (3.3 vs. 4.1). Hence, again the members of the former group – as they travel less frequently – earn fewer rewards. Similarly, also the share of peak trips (relative to off-peak trips) during the pre-measurement was higher among those who dropped out, which can be taken as another indication that it was relatively more difficult for this group to obtain the rewards for off-peak trips. This is confirmed by the observed share of peak trips during the reward period, which is substantially lower among participants who completed the experiment. Overall, the percentage of peak trips among those who completed the experiment dropped by 28.1% between the pre-measurement and the reward period, but only by 7.4% for those who dropped out. This gives a clear indication that participants dropped out because they did not manage to gain a sufficient number of rewards to compensate for the efforts related to the participation (filling in the logbooks and surveys, using the app, etc.). Finally, we find that participants with an iOS operating system had a higher propensity to complete the experiment. This is consistent with iOS users reporting to be more satisfied with the customized app than those who used Android-based smartphones (see Footnote 11).

Generally, we can thus conclude that – similar to the decision concerning participation in the experiment – also the decision to quit participation seems to have been made on rational grounds.

4. Describing travel behavior

4.1. Data processing

All participants had to install the customized app on their phone for the duration of the experiment, with a platform-specific version for Android and iOS. Overall the app has been installed on over 1000 phones, delivering over 8 million observations (timestamped geo-locations). The app registers the location of the smartphone in a central database. This happens when the smartphone detects a significant change in the location of the phone, upon which the new location is posted to the server using an active network connection. The server registers the location as well as the time of reception in the database. Additionally for each participant a customized list of up to 20 station locations (that are located between their origin and destination station) is registered in the app: when the phone detects the arrival or departure on one of these locations (using geo-fencing with a 100m radius), the app will post this event including the corresponding timestamped geo-location to the server.

Although the app drives the process of registration, the actual measurement of geo-locations as well as the communication with the database server happens at the level of the operating system and the hardware of the smartphone. Many variables can impact the registration of the observation, including the applied measurement method (GPS, wifi triangulation, etc.), the hardware and the availability and quality of a network connection. The collected timestamped geo-locations need

Table 4: Comparing those who participated in the entire experiment and those who quitted prematurely

	Participants who filled in all 6 logbooks	Participants who filled in 1–5 logbooks
Distance (reward) class		
≤ 25 km	20.5%	25.6%
25–40 km	25.8%	25.7%
> 40 km	53.7%	48.7%
Average number of weekly train trips during the pre-measurement	4.1	3.3
% peak trips during the pre-measurement	65.2%	74.5%
% peak trips during the reward period	46.9%	69.0%
%iOS (vs. Android)	46.1%	53.7%

to be further processed in order to remove noise as well as to identify the trips relevant for our analysis.[12]

In a first step we remove registrations from sequences that are temporally and spatially incompatible. Given that we study train travel, we assume a maximum speed of 160 km/h. Where registered speeds are higher, we analyze a series of subsequent registrations to assess which of them are least probable. In our assessment we account for smartphone-reported accuracy.[13] After removing this noise from the registrations we have for each participant the travel path during the entire experiment.

The second step is to identify distinct trips between the (participant-specific) origin and destination stations. We do this by defining participant-specific areas around the origin and destination stations. These areas have a radius of 20% (with a maximum of 5 km) of the geodesic distance between the origin and destination station. When the path identified in the previous step implies a trip between the two areas within a time limit of one hour plus an extra hour for each 60km, we use this observation as the starting point for the identification of the trip. To identify trip endpoints (which do not necessarily exactly coincide with station locations) we use a customized algorithm, which takes into account observed speed, distance to the origin and destination stations,

[12]Note that the algorithm discussed here is designed to analyze ex-post the choices made by the participants. It is not identical to the algorithm used during the experiment for the real-time allocation of rewards.

[13]Our algorithm to clean the observation dataset processes the point observations in reverse time order, starting with the last observation of the participant and working back to the first observation (excluding any observation outside the Netherlands). It searches for combinations of observations close in time for which the reported accuracy together with the corresponding geodesic distance imply an insignificant move (the sum of the squared accuracies exceeds the squared distance). If the implied speed for such a combination of two observations exceeds 160km/h, one of both observations is flagged as invalid based on a selection observations close in time (using spatial logic). Additional logic was added to handle observations with identical coordinates (but different timestamps). Such identical observations are caused by technical glitches, but also by the app reporting arrival at registered station locations (upon which it submits the station coordinates), or where the app bases location on a single radio source (submitting the coordinates of that source). Altogether our algorithm flagged 9.7% of the observations as invalid.

possible locations for transfers between trains, as well as the accuracy reported by the smartphone. For each trip we finally register departure and arrival time, and distance between the observed starting point and the origin station (and similarly for the end point). The average number of (valid) observations per trip is 12.8, with a median of 9.[14] The trips identified in this step may still include travel by modes other than train, as they could come no closer than 5km from the railway station that mark the begin and end point of the participation OD-pair, an area that may well cover home and work locations of the participant.

4.2. Selection of train commute trips

After following the steps described above, we can identify 97685 trips along the registered participation OD-pairs by 1182 participants. To identify *train commute trips* we make a selection using the following criteria:

- *Train* commute trips are defined as trips that have their begin and endpoint at the stations defining the participation OD-pair. We use a table of station locations defined by geographic coordinate and a radius that we extracted from the official timetable app. Smaller stations typically have a 200 m radius, whereas large stations are defined with a 525 m radius. Trips that begin and end at the registered station location are considered to be train trips.

- Only trips Monday through Friday with a departure time between 5.30am and 10am (for outbound trips) or between 15pm and 19.30pm (for return trips) are selected. The main motivation for limiting our analysis to this selection is our focus on scheduling preferences among train commuters. It is likely that for trips that take place during the rest of the day, these preferences are different (e.g. different preferred arrival times may apply). Moreover, the preferences related to travel attributes are usually found to be specific to the travel motive as well as the travel mode. Generally, the app does not identify the motive nor the mode. But with the recruitment targeting annual pass holders, it is reasonable to assume that by limiting our selection to peak trips on the participation OD-pairs we correctly identify commute trips.

- Registered trips that have taken place before or after the duration of the experimental period are removed.

- In the morning peak we only select trips in the home to work direction, whereas for the evening peak we limit to trips in the reverse direction.

These criteria result in a selection of 47215 train commute trips by 1009 participants.

The performance of our train commute identification algorithm can be validated against logbook registrations. Each logbook-registered trip that fits our definition of a train commute trip according to the above criteria, is matched (if available) against the corresponding app-registered trip (i.e. same day and same direction along the participation OD-pair). As Table 5 shows, the fit is rather good. The median absolute difference in departure time amounts to 3.68min and in arrival times to 3.20min. App-registered trips tend to start slightly later (technically there is a structural delay between the train departure signal which is registered by the participant in the logbook, and the train exceeding a speed threshold which is registered by the app; moreover participants may not register in the logbook small delays at departure), but end almost simultaneously (in terms of the median).

[14] For the selection of train commute trips discussed in section 4.2 the average is 14.3 and the median is 10.

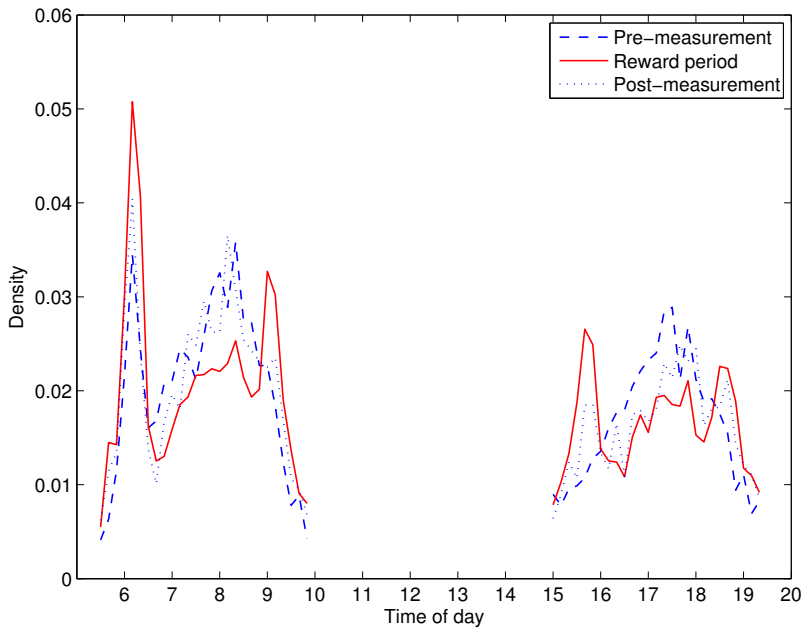
Table 5: Comparing logbook- and app-registered train commute trips in terms of departure and arrival time

	Median (in min)	Std. deviation (in min)
Departure time		
Absolute difference	3.68	32.14
App- minus logbook-registered time	2.04	33.54
Arrival time		
Absolute difference	3.20	32.50
App- minus logbook-registered time	0.01	34.14

4.3. Descriptives of travel behavior observed in the 3 phases of the experiment

Here we present descriptives of the set of train commute trips defined in Section 4.2. In Figure 1 we show the distribution of departure times over the time of day during the three experimental periods: pre-measurement, reward period, and post-measurement. The trips have been grouped into 10-minute intervals (e.g. between 7am and 7.10am) based on the observed departure time in the station where the trip originates. The surface under each of the curves is normalized such that each datapoint represents the share of a 10-minute interval in the total number of trips in the corresponding observation period.

Figure 1: Relative share of train commute trips by time of day (in terms of departure time)



As expected, we can observe a clear increase in the density of off-peak departure times during the reward period, while in the post-measurement period participants mostly return to their original behavior. But even during pre- and post-measurement, we observe peaks in the distribution of

departure times immediately before and after the peak hours. This pattern is likely to be caused by those participants who travel off-peak also without the presence of a reward being over-represented: clearly, train commuters who regularly travel outside the peak had a strong incentive to participate in the experiment, as they were able to collect the rewards without having to adapt their behavior (see also Section 3.1). But the observed pattern may also be a consequence of participants having changed their commuting arrangements in such a way that arrangements did not exactly coincide with the reward experiment.

In Table 6 we provide summary statistics across the three experimental periods. To calculate the trip frequency we divide the overall number of train commute trips by the number of participant-weeks; the latter is the number of unique combinations of week-number and participant ID in the train commute trip database. The assumption here is that if a participant is not observed traveling the whole week, this participant probably did suspend (or quit) his participation to the experiment for a variety of reasons (e.g. holiday).

Table 6: Changes in travel behavior over the experiment

	pre	reward	post	total
Nr. of trips per participant-week	3.75	3.82	3.75	3.80
Total number of trips	6789	35280	5146	47215
Morning peak	37.50%	29.24%	35.15%	31.07%
Evening peak	31.96%	24.14%	27.38%	25.62%
Off-peak	30.53%	46.61%	37.49%	43.31%
Decrease peak trips		21.76%	10.12%	
Decrease peak trips - high reward regime		22.60%		
Decrease peak trips - low reward regime		20.95%		

Table 6 demonstrates that trip frequency is almost constant across the three experimental periods, meaning that granting rewards for off-peak trips led to only few additional trips. The evolution of the relative share of trips during the three time periods (morning peak, evening peak, off-peak) in the number of trips clearly reflects the reward stimulus, with the off-peak share increasing from 30.53% in the pre-measurement to 46.61% during the experiment and decreasing again in the post-measurement period. The decrease in peak trip frequency (calculated again leaving out weeks in which a participant is not observed traveling) is 21.76%. This number is roughly the same for the morning and the evening peak. Moreover, the decrease is almost independent from the reward regime: it amounts to 22.60% when only trips during the high reward regime are taken into account, and to 20.95% when only trips during the low reward regime are considered.

Table 7 details for weekdays as well as for travel distances the relative share of peak trips as well as the decrease in peak trips relative to the pre-measurement. For the weekday-specific figures we do observe that the numbers are not wildly different, but we still note that the decrease in peak travel frequency as a consequence of the reward is the largest on Fridays, which happens to be the weekday when peak travel is relatively less intense on the train network. For Mondays we observe that in the post-measurement the decrease in trip frequency during the peak hours is bigger than

during the experiment; a dataset oddity that we could not find a good explanation for. As for the different distance classes we are not able to observe a clear pattern. While we do observe that members of the high distance class reduce the peak trip frequency more than members of the lower distance classes (supposedly due to the higher rewards granted to the members of the higher distance class), the difference is small and the ranking of the three classes inconsistent in this respect: the lowest decrease in peak trips is realized by the middle distance class.

Table 7: Changes in travel behavior over the experiment: weekday and distance class effects

		Share of peak trips (in %)			Decrease peak trips (in %)	
	Trips	pre	reward	post	reward vs. pre	post vs. pre
Weekday						
Monday	10357	68.47%	54.00%	64.50%	17.90%	23.24%
Tuesday	11057	69.03%	54.12%	63.21%	23.74%	3.29%
Wednesday	9110	69.55%	52.25%	62.93%	21.42%	1.33%
Thursday	10324	69.34%	53.99%	61.08%	21.24%	13.39%
Friday	6367	71.99%	51.76%	60.40%	25.71%	8.08%
Distance Class						
$<= 25$ km	10876	67.56%	53.85%	62.46%	23.04%	11.45%
$26 - 40$ km	11832	69.75%	55.96%	66.46%	12.09%	-0.59%
> 40 km	24507	70.21%	51.93%	60.65%	25.51%	14.33%

5. Modeling travel behavior

5.1. Discrete choice theory

We use discrete choice models to explain the observed scheduling behavior of the participants. In the models discussed below we focus on their choices between different train connections along their participation OD-pairs on a given day (considering days during the pre-measurement, the reward period and the post-measurement), for both the morning and the evening commute trip. These connections may not only differ in terms of departure time, but also in terms of rewards, travel time, schedule delays (relative to the preferred arrival or departure time), the number of transfers, crowding and unreliability. From the estimated relative preferences of the participants for the reward compared to the other attributes of the choice alternatives, their willingness-to-pay for reductions in travel time, schedule delays, the number of transfers, unreliability and crowdedness can be derived.

Discrete choice theory provides a broad range of modeling frameworks.[15] In general, it models

[15] An in-depth discussion on discrete choice theory can be found in Ben-Akiva and Lerman (1985); Train (1986); Anderson et al. (1992); Train (2003). The text in this section mainly draws on Knockaert (2010), Train (2003) and Bočkarjova et al. (2014).

the probability that a decision maker n chooses a given alternative j in choice situation m[16] as a function of the *random utility* U_{jmn} of the alternatives, which can be expressed as:

$$U_{jmn} = V_{jmn} + \epsilon_{jmn} \tag{1}$$

where:

- V_{jmn}: the *deterministic part* of the utility for alternative j as obtained by consumer n in choice situation m. We assume in our analyses that V_{jmn} is linear in parameters: $V_{jmn} = \beta x_{jmn}$ with β a vector of coefficients and x_{jmn} a vector of decision variables relating to consumer n and alternative j in choice situation m;
- ϵ_{jmn}: the *stochastic part*. The *multinomial logit* (MNL) model assumes that the stochastic part of the utility function follows a Gumbel distribution.

The consumer is then assumed to choose the alternative with the highest utility (utility maximization). In the MNL framework, the probability of choosing alternative j from choice set J in the choice situation m by decision maker n is given by:

$$P_{mn}(j) = \frac{e^{\beta x_{jmn}}}{\sum_{i=1}^{J} e^{\beta x_{imn}}} \tag{2}$$

Besides the MNL model, we will also estimate models where we allow for taste variation over respondents, meaning that the coefficients β in Equation (1) are distributed. We employ the *latent class* framework where the distribution of β is discrete (e.g. Greene and Hensher, 2003). Latent class models are a special case of mixed logit models (which generally assume a continuous distribution). In contrast to classical mixed logit models, latent class models do not require any distributional assumption – only the number of classes needs to be fixed in advance. Latent class models have the advantage that a closed form expression for the loglikelihood estimator is available, whereas mixed logit models require integrals to be calculated, and thus are computationally more demanding. As a consequence, it is rarely possible in mixed logit models to allow all relevant coefficients to be distributed, while in most latent class models this is feasible (as long as the number of classes is rather small, which is usually the case). Because of that as well as the presence of a limited number of classes, latent class models provide a rather insightful account of the heterogeneity present.

In a latent class model each class $c \in C$ has its own set of coefficients $\beta = b_c$ expressing conditional choice probabilities $P_{mn}(j|\beta = b_c)$ as in the multinomial choice model defined by Equation (1). On top of the class-specific choice models, a group membership model expresses the probability $P_n(c)$ that individual n is member of class c. A common approach, which we will also follow in this paper, is to use a simple logit model for group membership, as it allows for closed expressions for membership probabilities $P_n(c)$:

$$P_n(c) = \frac{e^{\gamma_c x_n}}{\sum_{i=1}^{C} e^{\gamma_i x_n}}, \tag{3}$$

where γ_c is a vector of coefficients for each class c defining the membership model and x_n is a vector of observed variables for person n. The coefficients $b_1, \ldots, b_C$ and $\gamma_1, \ldots, \gamma_C$ are estimated

[16] In our application choice situation m corresponds to a train commute trip along the participation OD-pair on a given day.

by identifying the βs that maximize the log-likelihood. In order to accommodate for correlations between repeated choices of travelers, latent class models with a panel structure (commonly referred to as 'panel latent class models') will be estimated:

$$LL = \sum_{n=1}^{N} \log \left[\sum_{c=1}^{C} \left(P_n(c) \prod_m \prod_j (P_{mn}(j|\beta = b_c))^{y_{jmn}} \right) \right] \quad (4)$$

where $y_{jmn} = 1$ when j is the chosen alternative by individual n in choice situation m, and $y_{jmn} = 0$ otherwise. Note that in the estimation the coefficients, γ_c must be normalized for one class.

5.2. Choice set

In order to analyze the participants' choice behavior we need to define a choice set for each selected trip. Choice sets consist of the relevant train connections along the relevant participation OD-pair (they thus differ across participants). Each choice alternative is characterized by the various attributes: reward, travel time, schedule delays, number of transfers, crowding, unreliability.

To define the choice set, we use timetable information from a popular smartphone app provided by the railway company (note that this a different app than the customized app used to observe the travel behavior of the participants). The app shows the relevant connections along a specific OD-pair. It is reasonable to assume that the app represents the relevant choice set for the participants, not least because all participants have a smartphone and presumably at least sometimes consult the app when selecting a travel connection. Since the choice set is assumed to be stable for each participant across the duration of the experiment, a more frequent (e.g. daily) check of the app is not necessary for this assumption to hold.

We designed a script to collect the train connections proposed by this app for all registered participation OD-pairs. We take into account all trips from home to work between 5.30am and 10am, and from work to home between 3pm and 7.30pm. The script was run for all weekdays at the end of the 2013 timetable period[17] (in the first two weeks of December 2013) to collect the information on the train connections. The information presented by the app includes departure and arrival times, the number of transfers, train types, as well as expected crowding.[18] Although the information was collected at end of the timetable period, we assume that it is representative for the entire period during which the experiment took place (August 2012-April 2013). However, as we will point out in Section 5.4, we will exclude observations along origin-destination-pairs where the match between the timetable and the observed trips is poor.

To define the choice set for each train commute observation (as defined in Section 4.2), we select the timetable-based departure time alternatives for the corresponding peak (morning or evening) of the corresponding weekday (Monday through Friday) of the week between Monday 2 and Friday 6 December 2013.

After defining the choice sets, we identify for each choice set the chosen alternative using the observed trip. We use the square sum of the difference in observed versus scheduled time of day at departure and arrival to define which of the timetable based connections was (most likely) travelled on.[19]

[17] The timetable period ran from mid December 2012 till mid December 2013.

[18] We will use some of this information to define the attributes of the choice alternatives, as described in Section 5.3.

[19] Visual inspection of the differences between the observed and logbook-reported departure and arrival times was used to evaluated this procedure. It yielded better matches than alternative procedures that were tested.

5.3. Utility function and attribute definition

We assume that participants make trade-offs between the reward, travel time, schedule delays, the number of transfers, unreliability and crowdedness when deciding on the departure time for the morning and evening commute trips. Our specification of the utility function can be interpreted as an extension of the modeling framework developed by Small (1982), which emphasizes the trade-off between travel time and schedule delays.

As for the choice set definition described in the previous section, we are able to extract information from the app provided by the railway company to define most of these attributes. An overview of the attributes, the notation used to describe them and the corresponding coefficients, and an account of how the attributes enter the utility function, is provided in Table 8.

A first choice attribute is the *reward*, denoted by R. During the reward period, it is positive for choice alternatives with a (planned) departure time outside the morning and evening peak periods. The size of the reward depends on the distance class that applies to a specific traveler and whether on the day associated with a specific choice situation he/she was in the high or the low reward regime (see Table 1). The reward equals 0 for all choice situations during the pre- and post-measurement.

A second attribute is *travel time*, TT, which we define as the interval between the (planned) departure time at the origin station of the trip and the (planned) arrival time at destination station of the trip. Both the planned departure and arrival times were extracted from the app in the same way as described in the previous section concerning the choice set definition.

Schedule delays generally describe the deviation between one's preferred schedule and the implied schedule when a specific alternative is chosen. More specifically, for morning commute trips we define them as the (absolute) difference between the planned arrival time and the preferred arrival time, and for evening commute trips as the (absolute) difference between the planned departure time and the preferred departure time. The underlying idea of this differentiation is that scheduling restrictions predominantly apply upon arrival for outbound trips, but upon departure for inbound trips. The preferred arrival and departure times required for the computation of the schedule delays are derived from survey outcomes.[20] Because of expected differences in the valuation of being early and late we specify separate scheduling attributes for earliness and lateness in the utility function (denoted by SDE and SDL, respectively). We also define them separately for the outbound morning (SDE^M, SDL^M) and the inbound evening commute trip (SDE^E, SDL^E), as the importance of schedule delays may differ between morning and evening commute.

Another attribute is the number of required *transfers*, TR, associated with a specific choice alternative. This information can also be extracted from the app provided by the railway company.

The *crowding* attribute, denoted by C, is based on the expected train occupation rate reported in the app. The app gives an indication of seat availability for most train connections (90.22% of the choice alternatives in the dataset have an indication of seat availability). Although the app visualizes the expected train occupation rate as a discrete variable with three levels, it actually receives the occupation rate as a continuous variable, which we will use as a proxy for comfort. In

[20] In the initial survey, participants were first asked at which time they would prefer to start and end work if they were able to travel anytime (hence, unconstrained by a timetable), comfortably and without any unexpected delays. Next we asked them what – conditional on their answers to the previous question – their preferred arrival time of the outbound train trip and their preferred departure times for the inbound train trip would be. Again they were explicitly asked to disregard existing timetables in their answers, as well as preferences of other travelers. Their answers are then used as reference points in the computation of the schedule delays.

Table 8: Overview of the attributes

Variable	Symbol	Units	Coefficient
Reward	R	€	β_R
Travel time	TT	hour	β_{TT}
Schedule delay early (morning)	SDE^M	hour	β^M_{SDE}
Schedule delay late (morning)	SDL^M	hour	β^M_{SDL}
Schedule delay early (evening)	SDE^E	hour	β^E_{SDE}
Schedule delay late (evening)	SDL^E	hour	β^E_{SDL}
Number of transfers	TR	#	β_{TR}
Crowdedness (as a proxy for comfort)	C	Expected train occupation rate in %	β_C
		Missing observation	β_{CU}
Unreliability	REL	% of trains delayed by more than 10 minutes at the destination	β_{REL}

the case where the connection is free of transfers, the crowding indicator value at the departure station (of the traveller) is used. When a connection implies transfers, the crowding indicator for each train is determined at the station where the traveller boards the train, finally the maximum value of train-specific crowding indicators is used as value for the whole connection.[21] Just as the choice set and the aforementioned attributes, the expected train occupation rates have been retrieved from the app during a specific week in December 2013. We assume that the data collected in these two weeks are representative for the entire duration of the experiment, meaning that the crowding attributes for a (departure-time-) specific train connection on a specific weekday is constant across the entire experiment. After testing different specifications of the crowding variable in the utility function, we decided to define crowdedness as a binary variable, specified as an expected occupation rate higher than 90% (this applies to 15.1% of the choice alternatives for which a comfort indicator is available). 90% seems to be the threshold where crowding begins to matter and not everyone might get a seat. Missing crowding observations are captured as an extra variable in order to avoid dropping observations.

Unreliability, indicated by REL, is defined as the chance that a traveler who chooses a specific train connection faces a delay of more than 10 minutes at the arrival station. It is the only choice attribute that cannot be retrieved from the app. Instead we use the logbook entries of the participants to compute a reliability indicator. The logbooks had to be filled in during 6 weeks in the course of the entire experiment, including the realized delays for all train commutes made during these weeks. We develop a model that predicts the chance of delays larger than 10 minutes at the arrival station for a specific origin-destination (OD) pair as a smooth function over the time-of-day. Note that due to limited data availability, the reliability indicator does not differ between days and train types. It only changes by the time of arrival at the destination station.

[21] We here apply the same methodology as the app does in showing one aggregated crowding indicator to the travelers.

The reliability indicator is based on a simple non-parametric function that weighs the delays[22] associated with all logbook observations of train trips that share the same arrival station as the OD pair under consideration.[23] Relatively higher weights are attached to trip observations that comprise the OD pair under consideration, and to those that have an arrival time close to the arrival time for which the reliability indicator is derived. The weights (bandwidths) are defined globally (hence they do not differ between OD pairs or between morning and evening commute trips) and are selected such that they minimize the prediction errors. As expected, we find that for most OD-pairs reliability is highest outside the peak periods. In the utility function, we specify reliability as a binary variable, which is specified as the chance of a delay (larger than 10 minutes) being lower or higher than 5%.[24]

Finally, the deterministic utility function related to consumer n, alternative j and choice situation m (see Eq. 1) is linear in parameters, and defined as follows:

$$\begin{aligned} V_{jmn} = \quad & \beta_R \, R_{jmn} + \beta_{TT} \, TT_{jmn} + \beta_{SDE}^{M,E} \, SDE_{jmn}^{M,E} + \beta_{SDL}^{M,E} \, SDL_{jmn}^{M,E} + \\ & \beta_{TR} \, TR_{jmn} + \beta_C \, (C_{jmn} > 90\%) + \\ & \beta_{CU} \, (C_{jmn} = \text{unknown}) + \beta_{REL} \, (REL_{jmn} > 5\%) \end{aligned} \tag{5}$$

5.4. Selection of observations for the discrete choice analysis

Before we proceed to estimate a choice model, there is a number of possible data quality issues that we want to consider. Based on the selection of the train commute trips described in Section 4.2, we will make a more narrow selection of observations that will be used in the discrete choice models:

1. The first issue is that some participants did not activate the app consistently.[25] For this reason we excluded all participants having no trip in either the pre-measurement period or the reward period (15.2% of the participants). We also excluded participants with a large difference in travel frequency between pre-measurement and reward period (absolute difference of more than three trips per week) (17.6% of the participants). The fact that the reward did not stimulate a change in travel frequency led us to believe that the observed behavior of these participants is caused by other unobserved variables that may, however, also affect scheduling behavior.
2. The second issue is related to the availability of the preferred arrival and departure times. The values of the independent variables $SDE_{jmn}^{M,E}$ and $SDL_{jmn}^{M,E}$ in the expression (5) for the deterministic utlitiy V_{jmn} are determined using the person-specific optimal arrival time at

[22] The delays are defined as dummy variable for all delays larger than 10 minutes at the arrival station.

[23] The condition that a common arrival station but no common departure station is required has been implemented due to a lack of observations that share the same departure and arrival stations on most OD-pairs. Clearly, it may render the reliability indicator less precise, but it ensures that reliability indicators are available for a large number of OD-pairs in our sample.

[24] 5% corresponds approximately to the median probability of delays exceeding 10 minutes in our sample. Alternative specifications and definitions of reliability were tested as well, yielding results that were rather similar to the ones presented here.

[25] Only 46% of the participants indicate that they had the app switched on during the entire experiment. The remaining participants state that they had the app switched on average during 77% of their peak-period travels and during 72% of their off-peak-peak-period travels.

work and departure time from work, which are available from the surveys. We therefore have to exclude from our analysis the observations from participants for which this survey information is not available (18.0% of the participants).

3. The third issue is related to the timetable information that we collected towards the end of 2013. While we are convinced that this information should be generally invariable over the entire timetable periods which started back in December 2012, there may have been small changes on individual origin destination pairs. Also may there have been shifts in allocation of train capacity impacting on the level of crowdedness. However, part of the experiment did take place during the previous timetable period. For a number of origin-destination pairs there was a significant change in train connections at the end of that period. To get an indication of the representativity of the timetable information we use the median square-sum of the matching of the observed trip to the timetable information. We calculate this median for each origin destination pair, for each timetable period, and separately for morning and evening peak trips. Where this median is larger than 0.025 (hours^2) we exclude the corresponding observations (9.7% of the trips).
4. The fourth issue is related to the observation of reliability. We exclude morning (evening) commute observations along OD-pairs where the number logbook observations at the arrival station is below 100 for the morning (evening) period (26.8% of the trips).
5. The final issue is related to weather circumstances during the experiment. For a number of days the weather forecasts made the railway operator decide to run a special winter timetable. We decided to exclude these days from the choice analysis (7.3% of the trips).

As a result, 22174 train commute trips are taken into account in the choice model estimations. It has been verified that the descriptive results on the evolvement of the relative share of peak trips over the three experimental periods (see Section 4.3) remains widely accurate also for this narrower selection of observations.[26]

5.5. Estimation procedure

In our dataset the number of trips is much larger than the number of respondents. The multinomial logit model assumes the error term ϵ to be independent across all observations and choice alternatives; an assumption that is likely to be violated when there are multiple observations by the same individual in the dataset. One way to deal with this issue is to accommodate for this correlation across choices by the same respondent in the specification of the utility formula. This is what we do when we specify a panel latent class model, where (probabilistic) class membership is modeled at the level of the participant rather than at the level of the choice.[27] But even in the case of the multinomial logit model we can accommodate to some extent for the panel structure of the dataset by considering the grouped robust estimation statistics. We refer to Freedman (2006) for a discussion of the topic.

In the estimation of the (panel) latent class models we observed that the (sub)optimum identified is dependent on the initial coefficient values. To find a global optimum loglikelihood we

[26]The relative share of off-peak trips for the narrower selection is 33.9% in the pre-measurement (compared to 30.5% in the wider selection), 47.6% in reward period (46.6% before), and 37.1% in the post-measurement (37.5% before).

[27]Note that in equation (4) the aggregation across the groups happens after the (conditional) aggregation across the observations for a respondent. This is equivalent to integrating over the (binomial) distribution of the coefficient vector β at the level of the respondent rather than the individual observation.

designed a random initialization procedure which allowed us to run a large number of estimations each with a different vector of initial coefficient values. The procedure is designed to automatically reinitiate model estimations (even in parallel on multiple CPUs) as long as a stopping criterium is not met. After each estimation the estimated coefficient values and the corresponding standard errors are compared to previous optima in order to decide whether a new (sub)optimum is identified.[28] The stopping criterium is defined as a heuristic that uses the number of model runs and the number of unique (sub)optima identified as independent variables. Estimation runs that fail to converge (as reported by the software or when the reported log likelihood (numerically) goes to minus infinity), that end in an area of numerical flatness (as reported by the software), or that result in a corner solution (in the estimation output one or more estimated coefficient values correspond to the boundary of the predefined estimation interval constraint) are not considered when evaluating the stopping criterium condition.

For the estimation of the discrete choice models presented in Section 5.6 we use a pre-release version of Pythonbiogeme 2.4 (Bierlaire, 2006).

5.6. Estimation results

Tables 9 and 10 give an overview of the estimation results and estimation statistics, respectively, for both the multinomial and the panel latent class logit model. Table 9 also shows the monetary valuations attached to the travel attributes. For the coefficients as well as the valuations, robust standard errors are presented.[29] The valuations can be obtained by dividing the attribute-specific coefficient by the negative of the marginal utility of income (i.e. the reward coefficient). They indicate how much on average the participants are willing to pay for improvements in the travel attributes (i.e. for a reduction in travel time, schedule delays, transfers, unreliability and crowdedness).

Looking at the results of the MNL model (as listed in the the left results column of Table 9), we find that all coefficients (and hence also the valuations) have the expected signs: a higher reward is associated with a higher utility, while increases in all other attributes have a negative effect on utility. All coefficients are significantly different from 0, except for the reliability and crowding coefficient as well as the coefficient associated with the missing observations for crowding.

Participants are on average willing to pay 15.5 Euro for reducing travel time by one hour.[30] This amount is somewhat higher than the official value used in Dutch cost benefit analyses concerning train travel for commuting purposes (11.5 Euro, see Warffemius (2013, p.16)). The difference may be attributable to the characteristics of our sample, which is not representative for train commuters in general due to self-selection and dropout decisions of participants (as described in Sections 2.2 and 3). Another reason explaining the divergence may be our use of revealed preference (RP) data, whereas the official value has been derived from a stated preference (SP) experiment. It is a common finding that travel time valuations are higher when RP data are used (e.g. Brownstone

[28] In a latent class setting the order of the (unlabeled) groups is not relevant for a behavioral analysis; we therefore standardize the model estimation output by ordering the classes arbitrarily before comparing the estimation to the (standardized version of) previous optima. In the procedure for identifying unique optima we use the classical standard errors; provided that we work with numerical results, any selection criterium is arbitrary so both the classical and the robust standard errors could be used here.

[29] For the valuations, the standard errors have been computed using the Delta method (e.g. Small, 2012).

[30] The result of a distinctively positive willingness to pay for reducing travel time contradicts research by Lyons et al. (2007). Based on a time use survey among train travelers, they suggest that in the face of the present ubiquitousness of portable electronic devices, train travel time *can and does possess a positive utility* [p.107].

and Small, 2005; Small et al., 2005), also in a public transport context (Wardman, 2001). Among the possible explanations are hypothetical biases and strategic answers in SP experiments, as well as travel time misperceptions (that affect SP and RP estimates differently).

A 1-hour reduction of the schedule delay is valued significantly lower than a 1-hour reduction in travel time. More specifically, the value of schedule delay early is 6.6 Euro/h during the morning commute and 5.0 Euro/h during the evening commute. The corresponding values for schedule delay late are 5.6 and 4.0 Euro/h, respectively. This implies that the participants have a higher disutility from deviating from their preferred schedules for the morning commute than for the evening commute. Moreover, they attach more disutility to being early than to being late, a pattern that has also been found in earlier peak avoidance experiments (e.g. Peer et al., 2015). The higher disutility from being early may capture that the valuation of earliness is not fully independent from the time of the day, which we do not explicitly take into account in our model (e.g. Tseng and Verhoef, 2008).[31]

Next we find that participants are on average willing to pay 2.8 Euro for reducing the number of transfers by one (note that the 2.8 Euro correspond to the disutility associated with ca. 11 minutes of extra travel time, which is a rather plausible value). This willingness-to-pay is likely to not only capture the actual penalty from changing trains, but may additionally capture that extra disutility associated with walking and waiting time, which tends to exceed the disutility associated with travel time spent inside the train (Wardman, 2001, 2004).

[31] During the early morning hours the utility of being at home tends to be rather high, thus rendering early arrivals at work particularly undesirable. Similarly, early departures from work in the afternoon may be undesirable, for instance because it may imply that one misses important meetings.

Table 9: Estimation results

		unit	multinomial logit		latent class logit					
					class 1		class 2		class 3	
			val	std err	val	std err	val	std err	val	std err
*Model coefficients**										
β_R	Reward	€	0.216	0.025	0.241	0.046	*0.0298*	0.0293	0.303	0.050
β_{TT}	Travel time	hour	−3.34	0.73	−5.00	0.85	−3.82	0.80	*−1.91*	1.04
β^M_{SDE}	Schedule delay early (morning)	hour	−1.43	0.11	−3.04	0.29	−4.14	0.30	−0.379	0.080
β^M_{SDL}	Schedule delay late (morning)	hour	−1.20	0.07	−3.80	0.22	−0.573	0.083	−1.27	0.13
β^E_{SDE}	Schedule delay early (evening)	hour	−1.07	0.08	−1.15	0.13	−1.46	0.13	−0.627	0.111
β^E_{SDL}	Schedule delay late (evening)	hour	−0.861	0.085	−2.01	0.18	−0.491	0.102	−0.411	0.148
β_{TR}	Transfer	#	−0.598	0.085	*−0.263*	0.157	−0.599	0.111	−0.941	0.173
β_C	Crowded train expected		*−0.0870*	0.0667	*−0.290*	0.167	*−0.0171*	0.0909	*−0.0256*	0.1460
β_{CU}	missing occupancy expectation		*0.0679*	0.1070	0.367	0.172	*−0.181*	0.154	*0.170*	0.201
β_{REL}	P(delay>10min)>5%		*−0.162*	0.129	*−0.00270*	0.20600	*−0.143*	0.149	*−0.357*	0.235
γ	Class constant						*0.197*	0.115	−0.354	0.135
Valuations†										
W_{TT}	Travel time	€/hour	15.5	3.3	20.7	4.2	(*128*)	122	(*6.30*)	3.25
W^M_{SDE}	Schedule delay early (morning)	€/hour	6.62	0.86	12.6	2.5	(*139*)	137	1.25	0.37
W^M_{SDL}	Schedule delay late (morning)	€/hour	5.56	0.65	15.8	3.0	(*19.2*)	18.8	4.19	0.93
W^E_{SDE}	Schedule delay early (evening)	€/hour	4.95	0.64	4.77	1.03	(*49.0*)	47.8	2.07	0.51
W^E_{SDL}	Schedule delay late (evening)	€/hour	3.99	0.56	8.34	1.49	(*16.5*)	15.4	1.36	0.56
W_{TR}	Transfer	€/#	2.77	0.56	(*1.09*)	0.70	(*20.1*)	20.5	3.11	0.81
W_C	Crowded train expected	€	(*0.403*)	0.313	(*1.20*)	0.81	(*0.574*)	3.147	(*0.0845*)	0.4788
W_{CU}	missing occupancy expectation	€	(*−0.314*)	0.489	*−1.52*	0.78	(*6.07*)	8.04	(*−0.561*)	0.657
W_{REL}	P(delay>10min)>5%	€	(*0.750*)	0.603	(*0.0112*)	0.8550	(*4.80*)	7.16	(*1.18*)	0.81
class share in sample			100.0%		34.2%		41.7%		24.0%	

*Coefficients that are not significant at the 5% level are in italics.

†Valuations that are not significantly different from 0 at the 5% level are in italics. Valuations for which either the corresponding model coefficient or the reward coefficient is insignificant are put into brackets.

Table 10: Estimation statistics

	multinomial logit	latent class logit
# observations	22174	22174
# individuals	544	544
LL_0	−64891.476	−64891.476
LL	−55471.579	−51258.689
estimation runs	11	1048
(sub)optima identified	1	91
no convergence	0	14

The value attached to reducing crowding (defined as expected train occupation rates of more than 90%) is positive, but close to 0 and not significant. Also when specifications with thresholds lower or higher than 90% were tested (including specifications with multiple thresholds), or when crowding was interacted with travel time, the coefficients remained insignificant. Our results suggest that the presence of crowding barely plays a role in departure time decisions. This finding contradicts most of the existing literature, which usually finds a significant negative effect of crowding on utility (e.g. Li and Hensher, 2011; Wardman and Whelan, 2011). A possible explanation again relates to our analysis being based on RP data, whereas the existing literature is almost exclusively based on SP data: in SP experiments, decision makers are fully informed about the (expected) crowding levels associated with the alternatives they can choose from. In real life, however, travelers may not be aware of the differences in (expected) crowding levels between different train connections. This may be due to a lack of experience with commuting during different times of the day. And even if travelers can revert to a vast travel experience, it may still be challenging for them to identify and remember the average crowding levels for different train connections, especially when the variation in crowding levels is low. A second explanation for the crowding coefficient being insignificant could be the quality of our data, which we collected from the app provided by the railway company, assuming that the expected level of crowding for a specific connection is constant over the course of the experiment. However, when comparing reported crowding levels (as indicated in the logbooks) and the expected crowding levels shown by the app, we find a fairly good match. This makes us confident that the expected crowding reported by the app is indeed an appropriate representation of actual crowding. Finally, correlation between the crowding variable and other attributes, as well as endogeneity, may play a role in explaining that the value of comfort is close to 0 in our sample.[32]

Similar arguments as for the crowding valuation apply also to the low and statistically insignificant willingness to pay for reductions in unreliability. Also when alternative definitions of unreliability or different utility specifications (e.g. including an interaction between unreliability and the OD-pair-specific train frequency) were employed, the reliability coefficient remained insignificant. In the existing literature on the valuation of reliability in train travel, a significant

[32]The comfort variable becomes endogenous if it is affected by the scheduling choices made by travelers. While it is unlikely that the participants of the experiment are a sufficiently large group to affect the comfort level (even more so, as we define comfort on a specific connection as fixed over the duration of the experiment), their choice behavior may still be fairly representative of all train commuters, meaning that comfort may be endogenous.

disutility associated with unreliable travel times is usually found (e.g. Bates et al., 2001; Warffemius, 2013).[33] But (again) the existing literature is predominantly based on SP data. And just as with crowding, travelers may not be aware of different reliability levels across train connections. Moreover, as described in Section 5.3, we were not able to identify the reliability of a specific train connection, but instead our reliability indicator is a continuous function over the time of the day, rendering our indicator possibly imprecise. This may have led to disutility associated with unreliability being picked up by other coefficients, such as the transfer coefficient. But apart from that, it should also be noted that the expected delays captured by our dummy variable (i.e. a probability higher than 5% to encounter a delay larger than 10 minutes) are rather low. If we make the assumption that an average delay longer than 10 minutes equals 20 minutes, the expected delay equals only $20 * 0.05 = 1$ minute. Also in earlier studies, the implied valuations for such small expected delays are very low. Here, longer expected delays (but then relative to the preferred rather than the expected schedule) are captured by the schedule delays. Our estimates for the willingness to pay for avoiding schedule delays imply that earliness and lateness do cause substantial disutility to travelers, and hence we expect that with more variation in our reliability attribute (which is typically present in SP experiments) we would have obtained a statistically significant value of reliability.

Next to the MNL model, we estimate a latent class model with 3 classes.[34] A 3-class model is chosen because it constitutes a good compromise between acceptable run times of the estimation procedure[35], reasonable significance of the coefficients, and still a rather detailed picture of the heterogeneity present in our sample. The table also shows that introducing the latent class structure leads to a substantial improvement in the loglikelihood compared to the MNL model. Overall, the results of the latent class model are quite consistent with the MNL results. All coefficients have the expected signs, and again all comfort and reliability coefficients are insignificant.

Class 1 captures preferences that imply rather high valuations of travel time and schedule delays. In contrast to the MNL estimation results, lateness is more costly than earliness. But in accordance with the MNL results, schedule delays concerning the morning commute are valued higher than those concerning the evening commute. The willingness to pay for reducing the number of transfers is small and not significantly different from 0 in this class.

The most remarkable characteristic of *Class 2* is the low and insignificant reward coefficient. This class, which accounts for a 42% share in the sample, thus captures scheduling behavior that is barely influenced by the presence of a reward. This is likely to hold true for those who traveled outside the peak already before the start of the experiment, which is a non-negligible number of participants in our experiment (see Figure 1 and Table 6). All valuations corresponding to this class are consequently insignificant as well (due to the reward coefficient being close to 0 they tend to be quite high in absolute terms). Moreover, Table 9 shows that in this class (just as in the MNL model), 1 hour of schedule delay early is preferred over 1 hour of schedule delay late for both the morning and the evening commute. The disutility associated with earliness is valued significantly higher during the morning commute, whereas for lateness the difference is not signifiant.

[33] Nevertheless, on the aggregate level the demand response due to changes in reliability tend to be small (e.g. Batley et al., 2011; van Loon et al., 2011; Preston et al., 2009).

[34] For reasons of simplicity we specify the class membership equation as multinomial logit model without explanatory variables.

[35] The estimation routine described in Section 5.5 required 1048 runs (resulting in 91 distinct sub-optima) for the latent class model presented here (see Table 10).

For *Class 3* we estimate a rather high reward coefficient, which in turn leads to fairly low valuations. For instance, the willingness to pay for reducing schedule delays by one hour ranges between 1.3 and 4.2 Euro. The travel time coefficient is insignificant in this class. The disutility from travel time might, however, be partially captured by the transfer coefficient, which is rather high in this class.

6. Conclusions

This study investigates how granting financial rewards to train commuters for traveling outside peak hours affects their departure time choices. We employ both descriptive analyses as well as random utility models that aim at explaining the observed choice behavior and at deriving monetary valuations for travel attributes. Our study is based on data collected during a large-scale, real-life experiment conducted in the Netherlands, which involved more than 1000 participants, who already had commuted by train before the start of the experiment. Depending on the reward regime and the distance class, they could earn between 1.5 and 4.5 Euro per off-peak trip along their participation OD-pair. Overall, the experiment lasted between 22 and 25 weeks, including 15-18 weeks of reward period. We measured the travel behavior of the participants using GPS observations from a customized smartphone app.

The reward experiment was successful in shifting trips from peak- to off-peak periods: the number of peak trips dropped by 21.76% among the participants as a result of rewarding off-peak travel. Interestingly, roughly half of this decrease in peak trips (11.12%) persisted during the post-measurement. These results give an indication that there is substantial potential for shifting train trips away from the peak period by charging fares that differ by time of day. Such time-differentiated fares can thus be a good alternative to typically very costly expansions of train and network capacity, and may even render it possible for train operators to reduce the overall supply of capacity (which is frequently determined by demand during peak hours).

We can explain the observed departure time choice behavior employing multinomial logit and latent class models. The estimation results can then be used to derive the participants' willingness-to-pay for reducing travel time, schedule delays, the number of transfers, crowdedness, and unreliability, using multinomial logit and latent class models. Our study forms an important contribution to the literature on train travelers' valuations of travel attributes. Unlike a large majority of existing studies, it is based on revealed preference (RP) (panel) data rather than stated preference (SP) data. While our valuations of travel time, schedule delays and transfers are fairly consistent with earlier findings, our result of insignificant valuations attached to improvements in reliability and comfort defers from the existing literature. Various explanations for the insignificance of the estimates can be brought forward, including little variation in the attribute values of the choice alternatives (and hence little effect on departure time choices), correlations in the attribute values, the travelers' lack of knowledge of the attribute values (especially during times with little travel experience), endogeneity, and issues related to data quality (we for instance make the assumption that the crowding and reliability attributes for a given train connection are constant across the entire experiment). Further research should be conducted in order to validate whether these findings can be generalized or whether they are a direct consequence of the setup of our experiment and the characteristics of the available data.

When interpreting the results presented in this paper, one should keep in mind that participants of the peak avoidance experiment were not selected randomly, but could self-select themselves. We showed here in a descriptive manner that participants are on average more flexible, more likely

to travel off-peak or at the edges of the peak period and also more likely to have a preferred arrival time outside the peak period compared to a surveyed group of non-participants. They can thus earn the rewards with less effort than non-participants. Due to the self-selection effect, we expect the decrease in peak trips of 22% to be biased upwards. Follow-up research will thus focus on modeling explicitly the decision to participate in the experiment, with the goal to obtain an indication what decrease in peak trips we can expect among the group of non-participants.

Besides an explicit model of the participation decision, future research that builds on the same dataset as this paper will mainly focus on extensions of the choice models. Here, the main focus was to demonstrate the application of the GPS data in multinomial and (panel) latent class discrete choice models with standard utility formulations. Future studies will focus on (1) joint modeling of the scheduling decisions concerning morning and evening commute (e.g. Jenelius et al., 2011), (2) a distinction between long-run and short-run scheduling decisions (e.g. Peer et al., 2015), possibly using days with a "winter timetable" (which have been excluded in this paper) for identification, as well as (3) enhancements of the utility specification, including the explicit consideration of socio-economic characteristics and time-of-day-dependent scheduling valuations (e.g. Tseng and Verhoef, 2008).

In our experiment we used a customized smartphone app to measure travel behavior. While representing an innovative technology (especially in the context of train travel), the app also imposed various challenges on the users (e.g. fast battery discharge, imprecise or missing registrations) as well as on our research using the resulting data. The fairly high number of participants dropping out before the end of the experiment might have been avoidable with a more user-friendly app. Evidently, future research using apps to measure travel behavior should put emphasis on developing and using a technologically advanced, yet user-friendly app. Additionally, interactive features of the app may be useful too: examples include the possibility to immediately check the travel registrations and revise them if wrong, as well as the possibility to enter additional trip information, e.g. on crowdedness and delays. Consequently, the functionality of the log-books (which had to be filled in on a separate webpage) could be integrated in the app. The benefits would be twofold: on the one hand a decrease in the necessary efforts of participation (and hence a lower dropout rate), and on the other hand a better quality of the information provided by the participants.

Future revealed preference experiments with voluntary participation should possibly also put more emphasis on the recruitment of the participants. Here, the recruitment via emails and posters in the main stations only induced few train commuters to join the experiment, likely augmenting the non-representativeness of the results. As an alternative, it may be sensible to approach train commuters (during peak hours) directly in the trains or at the stations in order to achieve a better response rate and a higher degree of representativeness. Finally, the response rate (and tenure rate) might also be raised by providing a fixed reward for participation (in addition to the departure-time-dependent rewards) in order to compensate for the efforts linked to participation.

References

Anderson, S. P., De Palma, A., Thisse, J. F., 1992. Discrete choice theory of product differentiation. MIT Press, Cambridge, MA.

Basu, D., Hunt, J. D., 2012. Valuing of attributes influencing the attractiveness of suburban train service in mumbai city: A stated preference approach. Transportation Research Part A: Policy and Practice 46 (9), 1465–1476. URL http://dx.doi.org/10.1016/j.tra.2012.05.010

Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. Transportation Research Part E: Logistics and Transportation Review 37 (2), 191–229.

Batley, R., Dargay, J., Wardman, M., 2011. The impact of lateness and reliability on passenger rail demand. Transportation Research Part E: Logistics and Transportation Review 47 (1), 61 – 72.
URL http://www.sciencedirect.com/science/article/pii/S136655451000075X

Ben-Akiva, M. E., Lerman, S. R., 1985. Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge, MA.

Bierlaire, M., 2006. BIOGEME: a free package for the estimation of discrete choice models. In: Swiss Transport Research Conference.

Bočkarjova, M., Rietveld, P., Knockaert, J., Steg, L., 2014. Dynamic consumer heterogeneity in electric vehicle adoption. Paper presented at the 2014 TRB Annual Meeting, Washington D.C.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies 17 (3), 285 – 297.
URL http://www.sciencedirect.com/science/article/pii/S0968090X08000909

Börjesson, M., 2008. Joint rp–sp data in a mixed logit analysis of trip timing decisions. Transportation Research Part E: Logistics and Transportation Review 44 (6), 1025 – 1038.
URL http://www.sciencedirect.com/science/article/pii/S1366554508000021

Broach, J., Dill, J., Gliebe, J., 2012. Where do cyclists ride? a route choice model developed with revealed preference GPS data. Transportation Research Part A: Policy and Practice 46 (10), 1730 – 1740.
URL http://www.sciencedirect.com/science/article/pii/S0965856412001164

Brownstone, D., Small, K. A., 2005. Valuing time and reliability: assessing the evidence from road pricing demonstrations. Transportation Research Part A: Policy and Practice 39 (4), 279–293.
URL http://dx.doi.org/10.1016/j.tra.2004.11.001

Carrion, C., Levinson, D., May 2012. Value of travel time reliability: A review of current evidence. Transportation Research Part A: Policy and Practice 46 (4), 720–741.
URL http://dx.doi.org/10.1016/j.tra.2012.01.003

Currie, G., 2010. Quick and effective solution to rail overcrowding. Transportation Research Record: Journal of the Transportation Research Board 2146, 35–42.
URL http://dx.doi.org/10.3141/2146-05

De Palma, A., Kilani, M., Proost, S., 2013. Discomfort in mass transit and its implication for scheduling and pricing.

Douglas, N., Karpouzis, G., 2006. Estimating the passenger cost of train overcrowding. In: 29th Australian Transport Research Forum.

Duncan, M. J., Badland, H. M., Mummery, W. K., 2009. Applying GPS to enhance understanding of transport-related physical activity. Journal of Science and Medicine in Sport 12 (5), 549–556.
URL http://www.sciencedirect.com/science/article/pii/S1440244008002107

Eliasson, J., Hultkrantz, L., Nerhagen, L., Rosqvist, L. S., 2009. The Stockholm congestion-charging trial 2006: Overview of effects. Transportation Research Part A: Policy and Practice 43 (3), 240–250.
URL http://www.sciencedirect.com/science/article/pii/S0965856408001572

Freedman, D. A., 2006. On the so-called "Huber sandwich estimator" and "robust standard errors". The American Statistician 60 (4).

Gong, H., Chen, C., Bialostozky, E., Lawson, C. T., 2012. A GPS/GIS method for travel mode detection in new york city. Computers, Environment and Urban Systems 36 (2), 131 – 139.
URL http://www.sciencedirect.com/science/article/pii/S0198971511000536

Greene, W. H., Hensher, D. A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. Transportation Research Part B: Methodological 37 (8), 681–698.
URL http://dx.doi.org/10.1016/S0191-2615(02)00046-2

Guo, Z., Wilson, N. H., 2011. Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. Transportation Research Part A: Policy and Practice 45 (2), 91 – 104.
URL http://www.sciencedirect.com/science/article/pii/S0965856410001564

Houston, D., Luong, T. T., Boarnet, M. G., 2014. Tracking daily travel; assessing discrepancies between gps-derived and self-reported travel patterns. Transportation Research Part C: Emerging Technologies 48, 97–108.
URL http://www.sciencedirect.com/science/article/pii/S0968090X14002290

Jenelius, E., Mattsson, L.-G., Levinson, D., 2011. Traveler delay costs and value of time with trip chains, flexible activity scheduling and information. Transportation Research Part B: Methodological 45 (5), 789–807.

Kang, B., Moudon, A. V., Hurvitz, P. M., Reichley, L., Saelens, B. E., 2013. Walking objectively measured: classifying accelerometer data with GPS and travel diaries. Medicine and science in sports and exercise 45 (7), 1419—1428.
URL http://europepmc.org/articles/PMC3674121

Karlström, A., Franklin, J. P., 2009. Behavioral adjustments and equity effects of congestion pricing: Analysis of morning commutes during the stockholm trial. Transportation Research Part A: Policy and Practice 43 (3), 283–296.
URL http://www.sciencedirect.com/science/article/pii/S0965856408001626

Kennisinstituut voor Mobiliteitsbeleid (KiM), 2013. Mobiliteitsbalans 2013. Tech. rep., Dutch Ministry of Infrastructure and Environment.

Knockaert, J., 2010. Economic and technical analysis of road transport emissions. Ph.D. thesis, Katholieke Universiteit Leuven.

Knockaert, J., Bakens, J., Ettema, D., Verhoef, E., 2011. Rewarding peak avoidance: the Dutch 'Spitsmijden' projects. In: Transitions towards sustainable mobility. Springer, pp. 101–118.

Knockaert, J., Ettema, D., Verhoef, E., Koster, P., Peer, S., Tseng, Y., 2012a. Spitsmijden Gouda-Zoetermeer [scientific report].
URL http://www.spitsmijden.nl/downloads/SM2D_scientific_report_final.pdf

Knockaert, J., Tseng, Y.-Y., Verhoef, E. T., Rouwendal, J., 2012b. The Spitsmijden experiment: A reward to battle congestion. Transport Policy 24, 260–272.
URL http://dx.doi.org/10.1016/j.tranpol.2012.07.007

Lam, T. C., Small, K. A., 2001. The value of time and reliability: measurement from a value pricing experiment. Transportation Research Part E: Logistics and Transportation Review 37 (2), 231–251.

Li, Z., Hensher, D. A., 2011. Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. Transport Policy 18 (6), 880–887.
URL http://dx.doi.org/10.1016/j.tranpol.2011.06.003

Lin, W.-H., Zeng, J., 1999. Experimental study of real-time bus arrival time prediction with GPS data. Transportation Research Record: Journal of the Transportation Research Board 1666 (1), 101–109.

Liu, R., Pendyala, R. M., Polzin, S., 1997. Assessment of intermodal transfer penalties using stated preference data. Transportation Research Record: Journal of the Transportation Research Board 1607 (1), 74–80.

Lyons, G., Jain, J., Holley, D., 2007. The use of travel time by rail passengers in great britain. Transportation Research Part A: Policy and Practice 41 (1), 107 – 120.
URL http://www.sciencedirect.com/science/article/pii/S0965856406000644

Mazloumi, E., Currie, G., Rose, G., 2010. Using GPS data to gain insight into public transport travel time variability. Journal of Transportation Engineering 136 (7), 623–631.
URL http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000126

Murakami, E., Wagner, D. P., 1999. Can using global positioning system (GPS) improve trip reporting? Transportation Research Part C: Emerging Technologies 7 (2–3), 149–165.
URL http://www.sciencedirect.com/science/article/pii/S0968090X99000170

Peer, S., Knockaert, J., Koster, P., Tseng, Y.-Y., Verhoef, E. T., 2013. Door-to-door travel times in rp departure time choice models: An approximation method using GPS data. Transportation Research Part B: Methodological 58, 134 – 150.
URL http://www.sciencedirect.com/science/article/pii/S0191261513001793

Peer, S., Verhoef, E., Knockaert, J., Koster, P., Tseng, Y.-Y., 2015. Long-run vs. short-run perspectives on consumer scheduling: Evidence from a revealed-preference experiment among peak-hour road commuters. International Economic Review 56 (1), 303–323.

Polydoropoulou, A., Ben-Akiva, M., 2001. Combined revealed and stated preference nested logit access and mode choice model for multiple mass transit technologies. Transportation Research Record: Journal of the Transportation Research Board 1771 (1), 38–45.

Preston, J., Wall, G., Batley, R., Ibáñez, J. N., Shires, J., 2009. Impact of delays on passenger train services. Transportation Research Record: Journal of the Transportation Research Board 2117 (1), 14–23.

Rietveld, P., 2002. Why railway passengers are more polluting in the peak than in the off-peak; environmental effects of capacity management by railway companies under conditions of fluctuating demand. Transportation Research Part D: Transport and Environment 7 (5), 347 – 356.
URL http://www.sciencedirect.com/science/article/pii/S1361920902000032

Rietveld, P., Bruinsma, F., van Vuuren, D. J., 2001. Coping with unreliability in public transport chains: A case study for netherlands. Transportation Research Part A: Policy and Practice 35 (6), 539–559.

Samenwerkingsverband Spitsmijden, 2009. Effecten van belonen in Spitsmijden in het OV: Hoe verleid je OV-reizigers?
URL http://www.spitsmijden.nl/resultaten/rresultaten2c/Rapport%20Spitsmijden%20in%20het%20OV%20in%20Spitsmijden%202%20aug2009.pdf

Santos, G., Shaffer, B., 2004. Preliminary results of the london congestion charging scheme. Public Works Management & Policy 9 (2), 164–181.
URL http://pwm.sagepub.com/content/9/2/164.abstract

Schönfelder, S., Axhausen, K. W., Antille, N., Bierlaire, M., Axhausen, K. W., Axhausen, K. W., Bierlaire, M., Bierlaire, M., 2002. Exploring the potentials of automatically collected GPS data for travel behaviour analysis: A Swedish data source. In: Möltgen, J., Wytzisk, A. (Eds.), GI-Technologien für Verkehr und Logistik. Vol. 13. Institut für Geoinformatik, Universität Münster, Münster, IfGIprints, pp. 155–179.

Schwanen, T., Dijst, M., Dieleman, F. M., 2002. A microlevel analysis of residential context and travel time. Environ. Plann. A 34 (8), 1487–1507.
URL http://dx.doi.org/10.1068/a34159

Shires, J., De Jong, G., 2009. An international meta-analysis of values of travel time savings. Evaluation and program planning 32 (4), 315–325.

Small, K. A., 1982. The scheduling of consumer activities: work trips. The American Economic Review 72 (3), 467–479.

Small, K. A., 2012. Valuation of travel time. Economics of Transportation 1 (1–2), 2 – 14.
URL http://www.sciencedirect.com/science/article/pii/S2212012212000093

Small, K. A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73 (4), 1367–1382.

Stopher, P. R., Jiang, Q., FitzGerald, C., 2005. Processing GPS data from travel surveys. In: 28th Australasian Transport Research Forum in Sydney, Australia.

Train, K., 1986. Qualitative choice analysis: Theory, econometrics, and an application to automobile demand. MIT Press, Cambridge, MA.

Train, K., 2003. Discrete choice methods with simulation. Cambridge University Press, Cambridge, UK.

Tseng, Y.-Y., Verhoef, E. T., 2008. Value of time by time of day: A stated-preference study. Transportation Research Part B: Methodological 42 (7-8), 607–618.
URL http://dx.doi.org/10.1016/j.trb.2007.12.001

van Loon, R., Rietveld, P., Brons, M., 2011. Travel-time reliability impacts on railway passenger demand: a revealed preference analysis. Journal of Transport Geography 19 (4), 917 – 925.
URL http://www.sciencedirect.com/science/article/pii/S0966692310001912

van Vuuren, D., 2002. Optimal pricing in railway passenger transport: theory and practice in The Netherlands. Transport Policy 9 (2), 95 – 106.
URL http://www.sciencedirect.com/science/article/pii/S0967070X02000057

Vickrey, W. S., 1969. Congestion theory and transport investment. American Economic Review 59 (2), 251–60.

Wardman, M., 2001. A review of British evidence on time and service quality valuations. Transportation Research Part E: Logistics and Transportation Review 37 (2–3), 107 – 128.
URL http://www.sciencedirect.com/science/article/pii/S1366554500000120

Wardman, M., 2004. Public transport values of time. Transport Policy 11 (4), 363 – 377.
URL http://www.sciencedirect.com/science/article/pii/S0967070X04000319

Wardman, M., Whelan, G., 2011. Twenty years of rail crowding valuation studies: Evidence and lessons from British experience. Transport Reviews 31 (3), 379–398.
URL http://dx.doi.org/10.1080/01441647.2010.519127

Warffemius, P., 2013. De maatschappelijke waarde van kortere en betrouwbaardere reistijden. Kennisinstituut voor Mobiliteitsbeleid (KiM).