Survey Research Methods (2025)
Vol. 19, No. 4, pp. 389-407
https://doi.org/10.18148/srm/2025.v19i3.8368
European Survey Research Association

The Effects of Study Duration on Nonresponse and Measurement Quality in a Smartphone App-Based Travel Diary

Danielle Remmerswaal[1] · Peter Lugtig[2] · Barry Schouten[2,3] · Bella Struminskaya[2]
[1]Utrecht University, Department of Methods and Statistics
[2]Utrecht University
[3]Statistics Netherlands

The use of smartphones for passive measurement can help increase response rates as such measurement may reduce the burden for participants by simplifying respondent tasks. Additionally, smartphone-based passive data collection can improve measurement quality by mitigating underreporting and reducing recall errors. In diary studies with reporting periods that last multiple days or weeks making these studies very burdensome, response rates are usually low and measurement errors can be substantial. In such burdensome studies smartphone sensor data can be particularly beneficial.

We report the results of a randomized experiment in an app-assisted travel diary study in a probability-based sample in the Netherlands. The app uses passively collected geolocation measurements to compile a diary consisting of trips and stops which respondents can edit and enrich. We invited half of the respondents to participate in the travel app for one-day, and the other half for seven-days (overall $N = 2544$). We also offered a one-day web diary as an alternative and varied the moment of offering the web diary.

More people registered in the app diary (12%) than in the web diary (7%). Unexpectedly, the initial app registration is significantly higher in the seven-day sample than in the one-day sample (11% vs. 14%). Study completion is significantly higher for the one-day participants (29% vs. 43%), making the absolute app completion rate the same (7%) for both groups. Using the population registry data, we analyzed whether socio-demographic variables related to travel behavior can predict participation in the app or web diary. We find differences in sample composition between the app and the web diary participants and between the one-day and the seven-day participants. Measurement quality in the app – measured with two dimensions – the amount of passive data collected and the active engagement of participants with their data – differs between the groups.

Keywords: travel diary; official statistics; reporting period; representativity; burden

1 Introduction

Smartphone apps have recently been tested as a method to (partially) replace traditional paper and online diaries (see e.g. Gillis, Lopez, and Gautama 2023; Jäckle et al. 2019; Fischer and Kleen 2021). Smartphone apps offer promising features to collect both digital behavioral data via sensors such as GPS and accelerometers, and self-report measures (Struminskaya & Keusch, 2020). The use of passive measurement on smartphones can be beneficial for response rates as it may reduce the burden for participants by replacing

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2025.v19i3.8368) contains supplementary material.

Corresponding author: Danielle Remmerswaal, Department of Methods and Statistics, Utrecht University, Padualaan 14, 3584CH Utrecht, Netherlands (Email: d.m.remmerswaal@uu.nl)

questions and simplifying tasks that respondents are asked to perform (Struminskaya et al., 2020; Toepoel et al., 2020). Additionally, sensor measurements can improve measurement quality by mitigating underreporting and reducing recall errors. In diary studies where response rates are usually low and measurement errors substantial, smartphone sensor data can be particularly beneficial. In such studies, respondents ideally report for multiple days or weeks, making diary studies very burdensome for respondents.

While there is a general agreement that respondents prefer shorter surveys over longer surveys (Revilla & Höhne, 2020), studies on the relationship between study duration and nonresponse produce mixed results. Galesic and Bosnjak (2009) found lower response rates and lower completion rates of longer web surveys for both the stated and actual length. Some studies find weak effects (Rolstad et al., 2011) or that other design choices, such as survey mode (see Cernat et al. 2022), have a larger effect on response rates. Dillman et al. (1993) find that shortening a questionnaire improves response rates but the effects differ among different populations. One reason for such mixed results can be that study duration represents only one factor contributing to response burden, which subsequently influences nonresponse and measurement quality.

In this paper, we report on a smartphone-based travel app study that uses passively collected geolocation measurements to replace questions. We invited 2,544 individuals from a probability-based sample in the Netherlands to participate in the travel app study over a one-day or seven-day period. The smartphone app automatically compiles a travel diary, consisting of a series of movements and stops, based on passively collected geolocation data (McCool et al., 2021). In this experiment, we test whether we can reduce burden of a travel diary in two ways: by offering a smartphone app next to a traditional online questionnaire, and by offering a shorter study duration. The mixed-mode design (concurrent vs. sequential) and the study duration of one-day versus seven-days were randomized (more information about the data collection and the experiments is given in the Data section).

Our focus in this paper is on the impact of the number of diary days indicated in the survey invitation on nonresponse and measurement quality. If a longer duration of an app-based diary study does not negatively impact response rates and measurement quality, this provides opportunities for more and richer diary data. Our research questions are: (1) how does the diary study duration (one vs. seven days) influence response rates, (2) how does the diary study duration (one vs. seven days) influence who participates in the app or the web mode of the study, and (3) how does the diary study duration (one vs. seven days) influence measurement quality measured as the amount of passively collected geolocation measurements and the level of active engagement in a travel diary app?

2 Background and Hypotheses

Response rates to surveys have been declining steadily over the past decades (Luiten et al., 2020). A survey type that requires special attention due to high nonresponse is the diary survey. For example, Household Budget Studies in the EU had a mean response rate of 50 % in 2015, with the Netherlands having the lowest response rate of 17% (Eurostat, 2020). Response rates of the Dutch travel survey ("ODiN") have been decreasing in the past decade from 50 to 25 % (McCool et al., 2021) despite increased efforts to reduce the response burden by moving the survey online and increasing the incentive.

Lower response rates can lead to less representative samples and if travel behavior of underrepresented groups differs from that of participants, also in a higher nonresponse bias. The level of nonresponse depends among others on the topic of the survey and the mode of data collection (Daikeler et al., 2020; Groves et al., 2000, 2004). In our study we aim to mitigate low response rates by offering a mobile app and a short study duration. We theorize that the relation between study duration and participation is mediated by study burden, that is, a shorter study reduces burden for participants, which then increases response and retention rates.

One of the first studies to investigate response burden is a study by Bradburn (1978) which stressed the importance of the interaction between the nature of the survey task (i.e. the objective burden), and the respondent's perception of the survey task (i.e. the subjective burden). Bradburn's conceptualization of burden involves four factors: survey length, frequency of being surveyed, required respondent effort, and the stress inflicted on the participant due to survey participation. Currently, survey researchers still differentiate between the *objective burden* and the *subjective burden* experienced by the respondent. Earp et al. (2022) used panel data to derive proxy measures for objective burden (number of months in sample), and for subjective burden (relevant household characteristics, and having item missing for stressful or sensitive questions). Read (2019) examined in a household budget app study the objective burden of respondents with the frequency and duration of app use, and subjective burden with self-report measures of how easy and interesting it was to participate. In a recent review on the concept of survey burden, Yan and Williams (2022) also reflect on the dynamic nature of survey burden by differentiating between the expected or *initial burden* at the time of a survey request and *cumulative* burden that respondents experience throughout the study. The initial burden influ-

ences response rates whereas the cumulative burden affects dropout rates and measurement quality.

For diary studies to ensure that a sufficient amount of data is collected, while at the same time minimizing dropout caused by high burden, researchers need to carefully consider the study duration. The advisable number of diary days differs per subject matter and the cycle in which activities of interest occur. Harmonized European Time Use Studies (HETUS) guidelines recommend to measure one weekday and one weekend day per respondent (European Commission. Statistical Office of the European Union., 2020). However, several Time Use researchers recommend to measure respondents for a week to account for intra-individual variation of behavior, since many activities follow a seven-day cycle (Glorieux & Minnen, 2009). For budget studies the study duration is commonly two weeks (Eurostat, 2020; Silberstein & Scott, 1991). The study duration of diaries used for medical research varies between a day and twelve months depending on the disease and the frequency of symptoms (Fischer & Kleen, 2021; Janssens et al., 2018). After testing the variability of travel behavior in multiple-week travel diaries, researchers recommend to measure for at least one week (Stanley et al., 2018; Schlich & Axhausen, 2003). Despite these recommendations, the common approach in travel data collection is to ask participants to keep track of their travels for one or two days in order to minimize response burden (Gillis et al., 2023).

Studies using apps for passive measurement are potentially less obtrusive for respondents, which allows us to conduct studies for longer periods (Toepoel et al., 2020). The idea is that by combining active and passive measurement on a smartphone, we can reduce the burden for respondents, although this depends on the study characteristics (Keusch & Conrad, 2022; Revilla, 2022; Toepoel et al., 2020).

In traditional cross-sectional surveys, participation and completion are often conceptualized as binary outcomes. However, in diary studies (te Braak et al., 2020) and studies involving passive data collection (Antoun & Wenz, 2021; Jäckle et al., 2019; Keusch et al., 2022) the participation process is more complex. In diary studies, participants need to familiarize themselves with the diary instrument and to repeatedly report on their behavior instead of only once. To participate in an app-based study respondents need to take some additional steps: downloading, installing, and registering the app are all part of the initial set-up. Besides, respondents may be asked to share additional data, which not everyone is able or willing to do (Keusch et al., 2019; Revilla et al., 2019; Struminskaya et al., 2021). These additional tasks add to the initial burden. Then, respondents need to continue using the app to complete the study task which adds to the cumulative burden.

Whether an app-based diary is memory-based (i.e., respondent input only such as entering expenses manually in a budget app), automated (i.e., no respondent input such as tracking geolocation), or semi-automated (i.e., respondent validates passively collected information in a diary) (Prelipcean et al., 2018) determines how much respondent input is necessary. There are large differences among semi-automated diaries. Harding et al. (2021) designed a burden spectrum for travel diary apps to help identify how burdensome a travel app is. The automated travel diary, which requires only passive logging, imposes the relatively lowest burden (see e.g. a travel diary by Patterson & Fitzsimmons (2016)). Giving more tasks to the respondent, such as an initial survey and validation tasks, increases both the objective and subjective burden. Technological advancements make fewer respondent tasks necessary. For example, when an app makes predictions of the transport mode based on the collected location data, the respondent does not have to supplement this information, but only has to accept or correct the predicted mode, which is less burdensome. Several studies have implemented transport mode prediction during the data collection. Supplementary Table A.1 provides an overview of recent smartphone-based travel studies that use passively collected geolocation data along with their key features.

A difficulty in assessing burden in app studies is that the relation between the time spent on the study and burden is ambiguous. Read (2019) found that in app studies, objective and subjective burden measures are not necessarily closely related. The length of the survey (number of questions or time spent) is a commonly used metric for objective survey burden. However, these metrics are unsuitable for measuring objective burden in app studies. While we can calculate the number of minutes participants spend in the app with paradata, that is, a by-product of the data collection process (Kreuter, 2013), this is not sufficient to estimate burden. The time spent might be more indicative of motivation than burden (Faghih Imani et al., 2020). Alternatively, in an evaluation survey, participants can be asked about their perceived burden.

2.1 Participation in App Studies

The study length is arguably the most obvious and intuitive factor of burden influencing survey participation. We know only of one app-based study using passive data collection that explored the effect of a reduced number of diary days on response rates, a travel app study by Svaboe et al. (2021). The one-diary diary had a slightly higher uptake than the earlier conducted seven-day diary, however, the design was non-experimental and the study duration was not the only alteration they made so the results are incomparable.

More research focused on the hypothetical willingness to participate in app studies using passive data collection.

A few studies in non-probability online panels have included study length as an experimentally varied factor. In vignette studies, respondents were 6–7 percentage points more willing to participate in a shorter one-month study than in a six months study (Keusch et al., 2019; Ságvári et al., 2021). Ságvári et al. (2021) found that the effect of study length was stronger amongst people with higher smartphone usage and fewer concerns about passive data collection. Ochoa (2022) tested a wider range of study durations (one week, 1, 3, 6, and 12 months) in a vignette study on sharing geolocation data. He found the same effect, participants prefer shorter study durations, and additionally found that the duration of the project is a more important attribute for deciding whether to participate than the amount of the incentive.

Most of the beforementioned studies are conducted in non-probability panels, only consider hypothetical willingness, and consider longer study duration than our study does. Nonetheless, we assume that by specifying the number of diary days in the invitation letter, we make the initial burden different for the two groups in favor of the shorter study duration. *We expect the app registration rates to be higher in the one-day group than in the seven-days group (H1.1).*

In app-based studies that collect measurements in the background, participants have to allow tracking on their devices. W*e expect no differences between the rates of the one-day and the seven-day group sharing geolocation measurements in the app (H1.2).*

In surveys in general, as well as specifically in diary surveys there is usually a large initial dropout followed by lower dropout rates over the course of the study (Lugtig, 2014; Tienda & Koffman, 2021; Hoerger, 2010; te Braak et al., 2020). Therefore, it is understandable that the actual length of the survey influences dropout, which is also found by Peytchev (2009). This trend is also visible in recent app-based diary studies, although technical issues on devices might play a role in enlarging the initial dropout (Jäckle et al., 2019; McCool et al., 2021). *We expect a higher completion rate in the one-day group due to the fact that their study task is shorter (H1.3).*

2.2 Representation in App Studies

Researchers warn about representativity issues in studies that collect data with smartphone apps due to people not being able or willing to participate (Keusch, Bähr, et al., 2020; Scherpenzeel, 2017; Wenz et al., 2019). Part of the explanation is coverage error, that is, the difference in smartphone access, while another part is in the use of the devices, or the (second-level) digital divide (Hargittai, 2002; Wenz & Keusch, 2022). People with low technical skills have lower levels of willingness to participate in studies that involve app and sensor-based data collection (Keusch, Struminskaya, et al., 2020; Wenz et al., 2019). This is supported by Jäckle et al. (2023) who found that one-third of invitees to their app study selected technical or ability related reasons for not participating.

Underrepresentation of certain groups becomes a problem when participants and nonparticipants systematically differ in the variables of interest (Peytcheva & Groves, 2009). Therefore, we investigate whether there are differences in demographic and travel-related characteristics of non-participants and participants in the app and the web mode of our travel diary study. *We expect differences in the composition of the samples that participate in the app vs. the web version (H2.1).*

Our main research question is whether certain groups are willing to participate for a longer period than others. To the best of our knowledge no studies have examined the relations between study duration and representativity specifically in the context of research apps with passive data collection. Previous research has identified a relationship between study duration and representativity in panel surveys. Revilla and Ochoa (2017) found that age negative influences the maximum desired length for a panel survey. Similarly, Revilla and Höhne (2020) reported negative effects of age on the desired and maximum survey length. Additionally, they found that a higher number of household members is associated with a reduced maximum desired length. *We expect differences in socio-demographic and travel-related variables between the participants in the one-day and the seven-day group (H2.2).*

Lugtig, Roth, and Schouten (2022) demonstrated that dropout from their app-based travel diary study did not affect selectivity. Te Braak et al. (2020) found that the dropout in their seven-day Time Use Study during the diary part was in the same direction as the first selectivity from the pre-questionnaire. Te Braak et al. (2020) find indications that people with busy schedules, measured with occupational status, dropout more often in time use diaries. This in contrast to Vercruyssen et al. (2014) who found that subjective busyness was related to participation and objective busyness was not. *We expect no differences in socio-demographic and travel-related variables between the participants who complete the app study and those who drop out (H2.3).*

2.3 Measurement Quality in App Studies

Data quality in diary studies is very dependent on the topic of the study. In Time Use Surveys, substantive measures can be used to evaluate quality in diaries (Chatzitheochari & Mylona, 2021). In household budget studies measurement errors such as rounding are used to assess data quality

(Berg et al., 2022; Jonker & Kosse, 2013; Schmidt, 2014). In travel surveys, we can assess measurement errors, for example by checking whether there are return-trips recorded. In travel app studies with passive tracking, there is an additional source that helps to assess measurement errors: geolocation measurements.

Several studies found that a longer survey length negatively affects data quality (Galesic & Bosnjak, 2009; Peytchev & Peytcheva, 2017; Toepoel & Lugtig, 2022). In multi-day diary studies, this effect is called the fatigue effect, where underreporting increases over the course of the study (Hu et al., 2017; Schmidt, 2014). In diary studies, however, there can also be a learning effect, so that provide more accurate information, and thus data quality increases over time. In our study, we can distinguish between the passive data collection by the smartphone sensors (i.e., where no actions besides, perhaps, activation are needed from the participant, such as geolocation measurement) and the active interaction of the respondent with the diary.

The passive data collection by smartphone sensors can be affected by dropout and phone malfunctioning. *On active days, we do not expect the number of geolocation measurements to be different between the one-day and the seven-day group (H3.1).* Since the geolocation measurement is performed continuously in the background, *we do not expect the number of geolocation measurements to be different between the first full day of the seven-day group and the mean day of the seven-day group (H3.2).*

The active engagement of respondents with the travel diary is more likely to be affected by learning and fatigue effects. However, the extent to which these processes operate similarly in app-based diary surveys using passive data collection, as opposed to traditional diary surveys, remains unclear. *We do not expect the engagement on the first day to be different between the one-day and the seven-day group (H3.3).* Although it can be a relatively low time consuming task to validate the travel diary (that is, to accept or correct the stops and tracks in the diary constructed based on the collected geolocations), we anticipate that over the course of the study, respondents will be less willing to complete and annotate the diary. *We expect the seven-day group to be interacting less with the diary in the app over time (H3.4).*

3 Methods

3.1 Sample, Experimental Design, and the Smartphone App

In this paper we report on a travel diary study that used a smartphone app to collect travel diary data over a one-day or seven-day period. The experiment was carried out between November 2022 and February 2023 by Statistics Netherlands (Schouten et al., 2024).

Respondents were recruited from the Dutch administrative register using a probability-based sample of the population in the Netherlands. The target population consists of non-institutionalized individuals aged 16 and older who live in the Netherlands. The survey was distributed by postal mail to a total of 3,211 individuals, consisting of two groups: 667 individuals were part of a follow-up sample who have previously completed a web-based Dutch National Travel Survey (ODiN) (CBS, 2023) and 2,544 individuals were a fresh sample. In the fresh sample, respondents were recruited using a cross-sectional sample of the population in the Netherlands drawn from the population register specifically for this study. The follow-up sample is analyzed in another paper (Klingwort et al., 2025) and will not be further discussed in this paper.

We tested two app versions: one with full editing functionalities and one with limited editing functionalities. Further elaboration on the app and the editing functionalities follows after the description of the recruitment strategy. We tested two recruitment strategies simultaneously: we varied the invited study duration and the mixed-mode design. The study duration of one-day (as it is used in the current web-based travel diary) was tested against a study duration of a week, seven days. The one-day group was instructed to use the app on a specified weekday. This group needed to download the app and register on the day before the specified day and use the app during 24 h on the assigned day. The seven-days group was instructed to use the app for seven days. This group was not instructed to start using the app on a specified day.

The second recruitment experiment concerns the mixed-mode design: we vary when the web questionnaire is offered as an alternative to app participation. The web questionnaire only had to be filled in for one day, but the day was assigned. The three moments of offering the web questionnaire were: (a) immediately mentioned in the invitation (i.e., a concurrent offer of the app option and web questionnaire option), (b) in the 1st reminder or (c) in the 2nd reminder (i.e., both sequential offers). Table 1 shows an overview of all experimental conditions with the number of invited samples and the realized response rates.

The incentive for app participants was a voucher worth 10 euros conditional on completion. The amount was the same for both study duration groups. Respondents who participated in the alternative web questionnaire version of the travel diary were only asked to provide information for one day. The web survey diary had a different incentive: a chance of winning an iPad or a voucher worth 400 euros.

The invitation letters were sent out by postal mail. The letter contained a unique username and password necessary to register in the app and/or web questionnaire. After regis-

Table 1

Overview of invited samples, realized samples, and experimental conditions of the entire study

Sample	Study duration app	Web mode in ...	Mode request	Incentive (€)	Invited samples: Follow-up Full edits	Invited samples: Fresh Full edits	Invited samples: Fresh Lim. edits	Response Rate (in %)[a] app	Response Rate (in %)[a] web
Follow-up	Seven days	Invitation	Both	20	667	–	–	32	21
Fresh	One day	Invitation	One	10 or lottery	–	212	212	10	9
Fresh	Seven days	Invitation	One	10 or lottery	–	212	212	11	11
Fresh	One day	1st reminder	One	10 or lottery	–	212	212	12	7
Fresh	Seven days	1st reminder	One	10 or lottery	–	212	212	13	7
Fresh	One day	2nd reminder	One	10 or lottery	–	212	212	10	4
Fresh	Seven days	2nd reminder	One	10 or lottery	–	212	212	14	4

The follow-up sample consists of participants of the 2022 web-based Dutch National Travel Survey. Regardless of the mode, the one day option is always accompanied by an assigned day. One mode request means respondent were asked to choose one of the diary options, both modes request means respondents were asked to complete the web diary and the app diary. In this paper, we do not analyze the follow-up sample due to the differences in sample composition and experimental set-up.
[a] AAPOR Response Rate RR2

tration, respondent's device information was recorded in the database. A short in-app introduction questionnaire (twelve personal and transportation-related questions) was the first thing participants encountered in the app. Secondly, participants were offered a short introduction tour in the app showing the app functionalities. Only then were participants asked to allow sharing of their geolocation data. The app started collecting geolocation data immediately after participants gave permission.

The app was compatible with both iOS and Android operating systems and was designed to work properly for phones not older than five years. Participants who had older models could still download the app and participate. The app sent no push notifications to avoid potential interactions with the study duration experiment. The app users who had been respondents in the app for at least 24-hours, for the one-day group, or five days, for the seven-day group, received their incentive voucher by postal mail alongside an invitation for an online evaluation survey of the app.

All the movements and stops collected by the app were visible to the respondent in a daily overview with a map (see Fig. 1a). By clicking on a movement recorded in the app (i.e., a trajectory with a starting and end point) a new screen opened on which respondents could edit their trip or delete the trip by clicking on the red bin symbol (see Fig. 1b). Participants were invited to validate each movement by labeling it with the mode of transportation used, and each stop by labeling it with the functional location (e.g., home or work). The functionality to label or delete movements and stops was available to all respondents. At the end of each diary day, respondents could submit the information about the entire day marking it as 'complete' and optionally enter a comment about their day. There was, however, no "submit" button, that is, no action from the respondent was needed to transfer data to Statistics Netherlands. The data was collected continuously and periodically sent to the servers when internet connection was available. The two additional editing functionalities in the app only available to the group with full editing options were: 1) modifying the starting and/or ending time of a movement or stop and 2) manually entering a movement or stop in the app.

Fig. 1c shows a screenshot of the overview of study days as visible to a participant. The one-day group was instructed that they could ignore days after the first full day and informed that they could continue for a full week, if they wanted to.

4 Analysis Plan

4.1 Nonresponse

For the overall app and web response rates we count the partial and complete responses (AAPOR RR2, (American Association for Public Opinion Research, 2023)). Then, for the app responses we make a subdivision based on the extent of participation. We use three stages of nonresponse based on the stages defined by Lugtig et al. (2022) in a study that used an earlier version of this travel diary app. Stage 1 indicates that a respondent downloaded and registered in the app. Stage 2 indicates participation, which we define as having provided at least one geolocation meas-

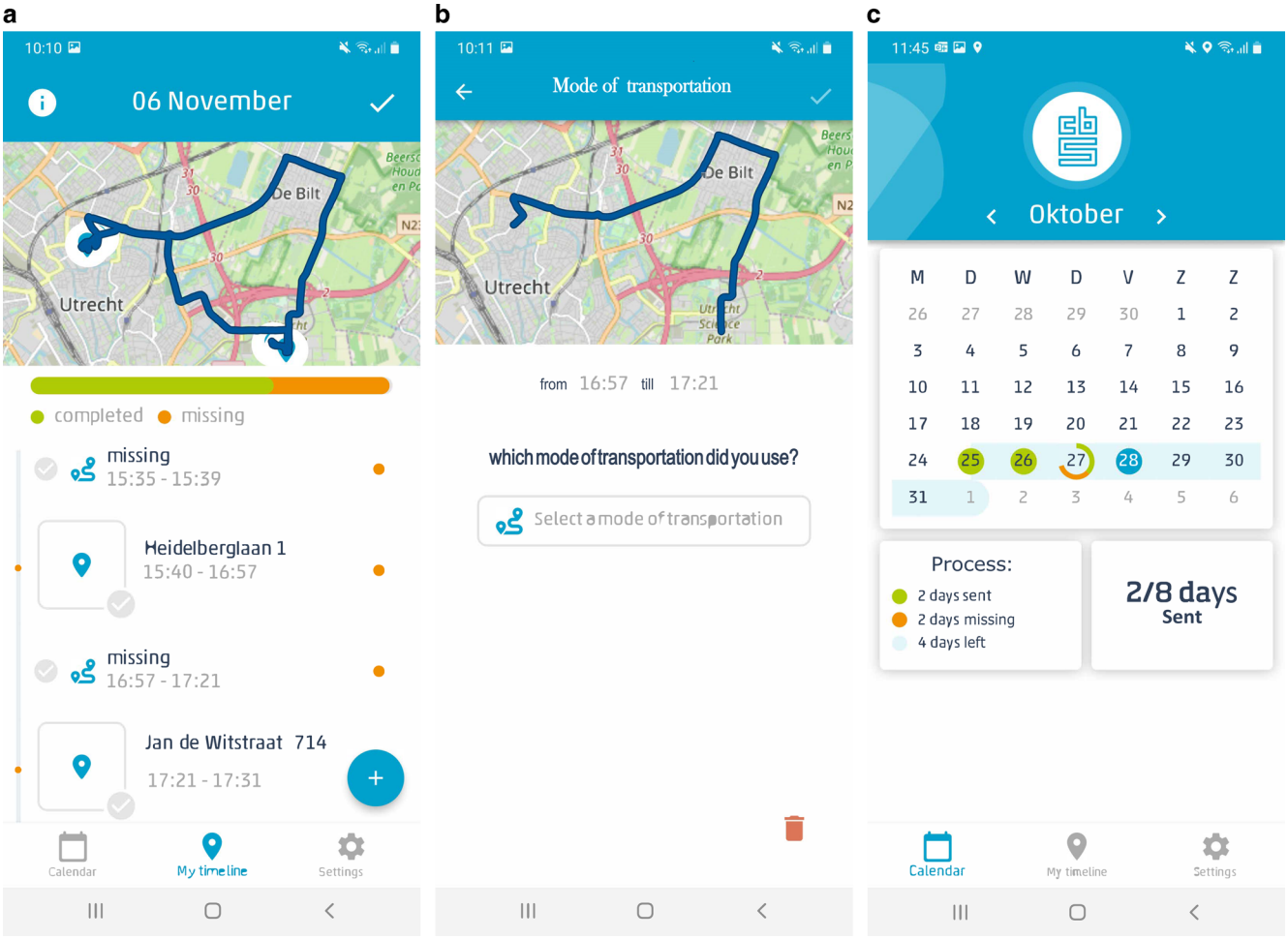

Fig. 1

***a** Daily overview **b** Validating a trip **c** Monthly overview with progress of participant*

urement. Stage 3 indicates study completion as described in the invitation letter, that is, keeping the app for one full day after the app installation for the one-day group and for seven days for the seven-day group. For each respondent, we determined whether the app remained installed based on the presence of either passive collection of geolocation data or active engagement of the respondents with the app (using app paradata). Additionally, we calculate the percentage of respondents reaching stage 2 conditional on stage 1, that is the percentage of the registered respondents that provides any geolocation measurements in the app, and the percentage reaching stage 3 conditional on stage 2, that is, the percentage of participants providing location measurements who complete the study (the definition of completion varies by assigned study duration).

For hypothesis 1.1, we compare the percentage in the one-day and the seven-day group reaching stage 1. For hypothesis 1.2, we compare the percentage in the one-day and the seven-day group reaching stage 2, and the percentage reaching stage 2 conditional on reaching stage 1. For hypothesis 1.3, we compare the percentage in the one-day and the seven-day group both for reaching stage 3, and for reaching stage 3 conditional on reaching stage 2. To assess the significance of the differences in the (conditional) response rates, we conduct χ^2-tests. To reflect the uncertainty in the data, we calculate Wilson 95% confidence intervals (CIs) around the proportions (Brown et al., 2001).

4.2 Representativity

We use a multinomial logistic regression model to evaluate hypothesis 2.1; we study whether there are any differences in the composition of the samples that participate in the app vs. in the web diary vs. do not participate in either. In the model we aim to predict participation in the app (stage 2—sharing geolocations) and in the web diary, while nonparticipation serves as the reference category. In-

dividuals who registered in both modes are treated as app respondents to avoid counting them twice, and to be able to include them in subsequent app analyses. We chose to focus on stage 2 participants since without sharing location data, their registration is not useful to Statistics Netherlands. We include socio-demographic respondent characteristics and the following study characteristics related to the recruitment strategy: study duration as stated in the invitation (one vs. seven days) and the moment of offering the web questionnaire (directly in invitation letter vs. in the first reminder letter vs. in the second reminder letter).

The socio-demographic variables we use to predict study participation were taken from the population register that is centrally available in the Netherlands, and therefore available for all invited sample members. We use the following variables that are predicted to be related to both participation and travel behavior (CBS 2023): gender (male vs. female), age classes (16–24 vs. 25–44 vs. 45–64 vs. 65+), migration background (born in Netherlands vs. not born in Netherlands), household disposable income classes (1st vs. 2d vs. 3th vs. 4th quartile), working status (employed vs. self-employed vs. unemployed vs. retired vs. other or unknown), household type (1 person vs. 2 (a couple) vs. 2+ (couple with children) vs. 1+ (other compositions including single parents)), degree of urbanicity (low (<500 addresses per km^2) vs. moderate (500–1500 addresses per km^2)) vs. high (>1500 addresses per km^2), car ownership (yes vs. no), driver's license (yes vs. no).

To compare the representativity of the sample participating in the app and web diary we calculate R-indicators and Coefficients of Variation (CV). For both these representativeness indicators we use the available socio-demographic variables in the model, so we exclude the study characteristics. Both representativity measures range between 0 and 1. Higher values of the R-indicator, and lower values of the CV indicate better representativity: that is, higher sample balance and a lower risk of nonresponse bias (for more details see Schouten et al., 2009; Shlomo et al., 2012). The CV is better suited for comparisons of samples with different sizes. Software code and a manual for the calculations of the R-indicator and the CV can be found at http://www.risq-project.eu.To reflect the uncertainty in the data, we calculate 95% confidence intervals around the R-indicators, and Coefficients of Variation. Confidence in intervals are calculated with the bootstrap method, using 1000 replications of the dataset, since an easy analytical derivation is not available.

To evaluate whether certain groups are willing to participate for a longer period than others (H2.2), we compare the characteristics of the individuals who register in the app or in the web questionnaire crossed with the study duration stated in the invitation. The multinomial logistic regression model with significant interaction effects can be found in Supplementary Table A.3.

We use a logistic regression model to analyze if socio-demographic or experimental variables predict whether participants who registered in the app complete the study (H2.3). We use the same socio-demographic variables as in the multinomial logistic regression models predicting participation and add the relevant experimental study characteristics, namely study duration as stated in the invitation (one vs. seven days) and available editing options (limited vs. extensive), and add variables describing the device used in the app, namely device type (Samsung vs. iPhone vs. other) and age category of device's operating system (OS) (new vs. old vs. neither new nor old) which we derived from paradata. We are not able to analyze potentially relevant respondent characteristics such as technological skills and privacy concerns since these are not available in the administrative register data.

4.3 Measurement quality

To answer the question whether the study duration as stated in the invitation letter influences the measurement quality, we assess the amount of data passively collected and the active engagement of participants with their data. Firstly, we consider the quantity of the passively collected geolocation measurements. We measure the quantity as the hours and minutes on a day with at least one sensor measurement collected.

Secondly, we consider the active interaction of respondents with the daily overview of their trips and stops. We measure this with the percentage of days in which respondents validate and label a trip or stop in the app diary, and by the percentage of days on which respondents with extensive editing options manually added a trip or a stop.

Additionally, we also consider technical issues with the app's processing of geolocation data into a diary overview with trips and stops. This was not the case for all participants on all days which hinders the ability to label trips and stops. To reflect the quality of the app processing we show the percentage of participants for whom trips and stops are compiled with geolocation measurements.

Of the eligible participants we compare the following five measurement quality outcomes per day:

- Mean minutes per day in which a geolocation sensor measurement was collected;
- Mean hours per day in which at least one sensor measurement was collected;
- Percentage of days with geolocation measurement for which trips and stops are compiled;

– Percentage of days with compiled diaries on which participants labeled any trip with a vehicle or any stop with a functional location (natural function of the place, e.g. work or home);
– Percentage of days with manually added stops or trips, for participants with extensive editing options.

To compare measurement quality between the one-day and the seven-day group (H3.1 and H3.3), we select the first full study day of all participants who registered. This is the day after registration, the day the one-day group is instructed to start. By selecting the first full day instead of the registration day, we omit a learning period and make sure that everyone has potentially twenty-four hours to send geolocation measurements. Additionally, we compare these days to the mean day of the seven-day group, to check for potential changes in measurement quality over time (H3.2 and H3.4). We construct 95 % confidence intervals around all estimates with the non-parametric bootstrap method using 1000 replications of the dataset. The variables on the amount of sensor measurement per day have a distribution with bounded support, they can take values between 0–24 h and thus 0–1440 min. Because the variables are not normally distributed we used the distribution-independent bootstrap method for calculating the 95 % confidence intervals instead of a standard closed-form formula (Efron & Tibshirani, 1993).

All analyses were run in R (version 4.2.3) (R Core Team, 2021). The code is available in the replication materials. The datafiles loaded in the R code are property of CBS (Statistics Netherlands) and contain micro-data from the population register, and can therefore not be openly shared.

5 Results

5.1 Stages of Nonresponse

In the fresh sample, 2,544 persons were invited to the study, of which 315 registered in the app (12 %, AAPOR RR 2), and 178 completed the questionnaire (7%, AAPOR RR 2). Although they were asked in the invitation letter to participate in one mode, 10 respondents both registered in the app and completed the questionnaire. A total of 483 respondents participated (19 %, AAPOR RR 2). In the remainder of this section we will focus on the app respondents.

As explained in the analysis plan, we use three stages to analyze response behavior of participants in the app and answer the question whether people are more likely to participate in an app-based travel diary for one or seven days. The (conditional) response rates for each stage are shown in Table 2. Contrary to our expectations, the initial registration is lower in the one-day group (11 %) than in the seven-day group (14 %). The difference is statistically significant ($\chi 2(1) = 4.70$; $p = 0.03$) and we reject H1.1. Most people who registered in the app reached stage 2 and sent geolocation measurements. The percentages reaching stage 2, and the percentages reaching stage 2 conditional on stage 1, do not significantly differ between the one-day group and the seven-day group thus we reject H1.2. Note that the difference between the two groups reaching stage 2 is not significant anymore, while it was for stage 1. Not everyone completed the study and reached stage 3. A sizeable amount of 29 % of the one-day group dropped out within 24-hours and 43% in the seven-day group dropped out before the 7th day. This means that a significantly higher percentage of the participants in the one-day group completed their instructed

Table 2

Response rates (RRs) travel app per experimental treatment group

	One-day group				Seven-days group				Contrasts	
			95% C.I.				95% C.I.			
RRs per survey stage	*n*	%	Lower	Upper	*n*	%	Lower	Upper	X^2	*p*-value
Stage 1: registration	139	11	9	13	176	14	12	16	4.70	0.03
Stage 2: participation	133	11	9	12	159	13	11	14	2.61	0.11
Stage 3: completion	94	7	6	9	92	7	6	9	0.01	0.94
Stage 2 conditional on Stage 1	–	96	91	98	–	90	85	94	2.04	0.15
Stage 3 conditional on Stage 2	–	71	62	78	–	58	50	65	4.89	0.03
N invited	–	1272			–	1272			–	–

The table shows the response rates per (conditional) survey stage for the two experimental treatment groups that were invited to participate for one-day or for seven-days in the travel app. *n* denotes the number of respondents. Wilson 95% Confidence Intervals (C.I.) were used. We conducted pairwise contrasts of percentages

study duration ($\chi 2(1) = 4.89$; $p = 0.03$) and we thus do not reject H1.3.

5.2 Representativity

We evaluate representativity in three steps. First we present the results of a multinomial regression for web and app participation as a basis for the representativity indicators.

Table 3

Multinomial logistic regression results of app and web participation

	Non participation		App participation		Web participation	
	AME	Std. Error	AME	Std. Error	AME	Std. Error
Study duration in invitation (ref. 1 day)						
7 days	–0.031*	0.015	0.021	0.012	0.010	0.009
Moment web diary (ref. in invitation letter)						
In 1st reminder letter	0.011	0.018	0.014	0.015	–0.025*	0.012
In 2 d reminder letter	0.035*	0.018	0.012	0.015	–0.047**	0.012
Sex (ref. Male)						
Female	–0.011	0.015	0.005	0.012	0.006	0.010
Age: (ref. 16–24 years)						
25–44	0.028	0.027	–0.021	0.024	–0.007	0.017
45–65	0.040	0.028	–0.042	0.024	0.002	0.018
65+	0.031	0.038	–0.051	0.031	0.020	0.025
Born in NL (ref. no)						
Yes	–0.041*	0.018	0.046**	0.014	–0.005	0.013
Income percentile (ref. 1–24)						
25–49	0.026	0.022	–0.006	0.019	–0.020	0.013
50–74	–0.053*	0.025	0.032	0.020	0.020	0.016
75–100	–0.091**	0.027	0.065**	0.022	0.026	0.017
Working status (ref. employed)						
Self-employed	0.042	0.024	–0.046*	0.019	0.004	0.017
Unemployed	–0.005	0.038	–0.009	0.032	0.014	0.024
Retired	–0.028	0.034	–0.013	0.028	0.041	0.023
Other or unknown	0.041	0.041	–0.041	0.035	0.000	0.026
Household type (ref. 1 person)						
2	–0.073**	0.021	0.040*	0.018	0.032	0.013
2+	–0.033	0.022	0.018	0.018	0.015	0.014
1+and other	0.005	0.029	0.014	0.026	–0.019	0.016
Urbanicity (ref. low)						
Medium	0.018	0.017	0.006	0.018	0.016	0.015
High	–0.021	0.022	–0.003	0.014	–0.015	0.011
Driver's license (ref. no)						
Yes	–0.019	0.022	0.018	0.017	0.001	0.015
Owning a car (ref. no)						
Yes	–0.010	0.017	–0.006	0.014	0.016	0.011
N				2544		

N = sample size
* $p < 0.05$, ** $p < 0.01$.

We enter all auxiliary variables at once and do not perform variable selection. Second, we explore the role of interactions between the experimental condition study duration with various auxiliary variables. Here, we do perform a variable selection procedure. Finally, we evaluate completion through a logistic regression. Here, again we enter all relevant variables into the model.

5.3 Explaining Web and App Response

Table 3 shows the results of the multinomial logistic regression that compares background characteristics of three groups: participants in the app who reached stage 2, web diary survey participants, and non-participants.

We find significant effects of the recruitment strategies on study participation. People invited for a study duration of one day are 3 percentage points (p.p.) less likely than those invited for seven days to participate in the study at all. The study duration effect is not significant for the app or web mode separately. The timing of offering the web questionnaire (directly or in a reminder letter) influences the response rates; the probability of participation in the web diary decreases when the web mode is not offered directly (offered in 1st reminder: –2.5 p.p., offered in 2d reminder: –4.7 p.p.). Interestingly, offering the web questionnaire later had no significant effect on response rates of the app.

On basis of the available socio-demographic variables we calculated representativity measures for the app and the web respondents. The 95 % confidence intervals of the R-indicators overlap, meaning that the overall representativity of the app (R-indicator = 0.90 [95 % C.I. 0.88; 0.93]) and the web (R-indicator = 0.92 [95 % C.I. 0.89; 0.94]) do not differ significantly. However, the Coefficient of Variation, which takes the sample size into account, is worse for the web (CV = 0.50 [0.41, 0.60]) than the app (CV = 0.35 [0.29, 0.41]). The combined response of the app and web diary results in slightly improved coefficient of variation (CV = 0.32 (95 % C.I.: 0.29; 0.38)) and a slightly worse R-indicator (R-indicator = 0.88 (95 % C.I.: 0.893; 0.946)).

We do find significant effects on participation of the following socio-demographic variables: income class, working status, migration background, and household type, thus we do not reject H2.1. Those in the highest income class are 6.5 p.p. more likely to participate in the app, compared to those in the lowest income class. People in the second income quartile are 5.3 p.p. more likely to participate in the study (any mode), compared to those in the lowest income class. Self-employed are 4.6 p.p. less likely than those who are employed to participate in the app. We find similar effects as the previous travel app study did for migration background (Lugtig et al., 2022). Individuals born in the Netherlands are significantly more likely to participate in the app compared to those not born in the Netherlands. We do not find such an effect for the web diary.

Consistent with the findings of the previous travel app study by Statistics Netherlands (Lugtig et al., 2022) is that owning a driver's license and owning a car, which are strongly related to travel behavior, are both not related to participation behavior. Members of two-person households are 7.3 p.p. more likely to participate in the study than not, compared to people living alone. The results of the multinomial logistic regression with the significant interactions of variables with study duration can be seen in Supplementary Table A.3. The results imply that one-person households are less likely to participate for seven days than for one day. We added interactions of various auxiliary variables with study duration to the multinomial logistic regression. We found only household type to have a significant interaction effect with study duration. Thus, we do not reject H2.2. The results of the multinomial logistic regression with this significant interactions of household type with study duration can be found in the Supplementary (Table A.2). The results imply that one-person households are less likely to participate for seven days than for one day.

5.4 Explaining completion

Table 4 shows the outcomes of the logistic regression model predicting who completes their study after app registration. Respondents who complete the study (as defined by their respective study durations) are less frequently living alone. Other socio-demographic and travel-related variables show no significant effects so we do not reject H2.3. We do find significant effects on study and device characteristics. Participants registering with a smartphone other than an iPhone or a Samsung are 22.3 p.p. less likely to complete the study. The age of the operating system (OS) does not influence dropout rates. Participants with limited editing capabilities in the app are 14.5 p.p. more likely to dropout than those with extensive editing options. Lastly, participants invited for a seven-day diary are significantly less likely to complete the study, as also visible in Table 2 with the participation rates.

5.5 Measurement quality

We first summarize the results for the announced data collection period, the first seven days in the app. By our design, respondents in the one-day group only had to participate for one full day after which they were allowed to delete the app, but if the app was still on their phone it would continue measuring for the entire period of seven days. To

Table 4

Logistic regression results of study completion for app participants

Reference = no completion	Completion	
	AME	Std. Error
Study duration in invitation (ref. 1 day)		
7 days	–0.142**	0.055
Sex (ref. Male)		
Female	–0.034	0.056
Age: (ref. 16–24 years)		
25–44	0.108	0.095
45–65	0.061	0.106
65+	–0.024	0.127
Born in NL (ref. no)		
Yes	0.097	0.082
Income percentile (ref. 1–24)		
25–49	0.077	0.110
50–74	–0.003	0.098
75–100	0.052	0.096
Household type (ref. 1 person)		
2	0.223*	0.089
2+	0.133	0.093
1+and other	0.256*	0.125
Urbanicity (ref. low)		
Moderate	0.001	0.081
High	0.048	0.065
Driver's license (ref. no)		
Yes	–0.101	0.082
Owning a car (ref. no)		
Yes	0.059	0.064
Device (ref. iPhone)		
Samsung	–0.040	0.063
Other	–0.223*	0.099
OS age (ref. new)		
Average	0.063	0.067
Old	–0.028	0.099
Editing options (ref. extensive editing)		
Limited editing	–0.145**	0.056
N	315	

N number of app registrants (stage 1)
Working status groups are omitted, no self-employed completed the study.
* $p < 0.05$, ** $p < 0.01$

compare the measurement quality between the one-day and the seven-day group we concentrate on the first full day of both groups. We also compare this to the mean day of the seven-day group as to investigate whether measurement quality changes over the week for the seven-day group. The results are shown in Table 5.

In the app, geolocation data was collected for 133 respondents in the one-day group and 159 in the seven-day group—that is the participants that reached stage 2. The mean number of days for which geolocation data was collected is higher for the seven-day group (5.7 days) than for the one-day group (4.5 days). This means that many respondents provide data on more diary days than we asked them to, which is in line with other app-based travel diary studies (Gillis et al., 2023).

On the first full day geolocation data was collected for 95 respondents in the one-day group and 115 in the seven-day group. Both groups have on average 16 h with geolocation data on the first full day. Thus, we do not reject H3.1. Over time, the seven-day group provides less hours with location measurements as visible by the lower average amount of hours and minutes with geolocation measurements on the first full six days of the seven-day group. This is different than we expected (H3.2). However, for all groups the variation of the amount of hours with geolocation measurements is high, some participants provide geolocation measurements on only one hour while others provide data for the full twenty-four hours of a day.

For some participants the app did not compile a diary for some days on which they had collected geolocation measurements. On those days, participants were not able to see trips and stops in their daily overview in the app and thus also not able to validate them. We speculate this is either because respondents did not open the app anymore, which was necessary for the app to compile the data into a diary, or because of technical issues of the app. For measures concerning the active engagement with the app we consider the participants who were able to see trips and stops and interact with them in their daily overview in the app. Contrary to our expectation (H3.3), a higher percentage in the one-day group (79 %) labeled any period than in the seven-day group (73 %). Among the respondents with extensive editing options who reached stage 2, roughly the same percentage of participants in both the one-day and the seven-day group used the function of adding trips and stops manually on their first day (23 % vs. 20 %). Over time, the active engagement with the app did not change much for the seven-day group so we reject H3.4.

Noteworthy, the travel app did not perform equally well for all types of devices and brands. Older smartphones and certain brands registered fewer geolocation points and had more missing periods. The differences between iOS and Android were, however, relatively small.

Table 5

Measurement quality of users with any geolocation data collected on first full day

	One-day group								
		95 % C.I.			95 % C.I.			95 % C.I.	
	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper	Lower
Mean minutes per day with a geolocation measurement	661	565	763	608	520	700	502	459	540
Mean hours per day with at least one geolocation measurement	16	14	17	16	14	17	13	12	14
Percentage of days with compiled trips and stops	88	84	93	84	79	89	89	86	92
Percentage of compiled diary days with labels	78	71	85	73	66	80	69	65	73
Percentage of days with manually added stops or trips (participants with full editing only)	23	14	32	20	12	28	25	20	30
Diary days	95			115			538		

95 % Confidence Intervals (C.I.) are based on 1000 bootstrapped datasets

6 Discussion

In this paper, we reported on an app-assisted travel study that collected geolocation measurements to compile a diary consisting of trips and stops which respondents could edit and enrich. We conducted a randomized experiment in which 2544 individuals from a probability-based sample in the Netherlands were assigned to either participate in the travel app for one day or for seven days. As an alternative mode to the app, we offered a one-day web diary, either directly in the invitation letter or in one of the two reminder letters. We varied the study length in order to learn about the respondent trade-offs between active engagement, burden and the logic of employing a mobile app. We varied the timing of the web questionnaire in order to learn about respondent preferences in mode (smart app versus 'non-smart' web questionnaire). In our study design, respondents were not aware of the experimental conditions. Our research questions address the impact of diary study duration on rates and representation of registration and engagement, and on measurement quality. We do not explicitly report the impact of the timing of the web questionnaire invitation, but we do compare representation between the modes.

We find that diary study duration has a modest impact on registration, measurement quality and the choice of app versus web questionnaire, but that drop-out can be sizeable. However, offering the alternative web questionnaire sequentially instead of concurrently did not significantly increase the response rates for the app diary but did significantly reduce the response rate of the web diary. Also, the modes differ in which respondents they attract., and, hence, representation.

Contrary perhaps to expectations and vignette studies findings (Keusch et al., 2019; Ochoa, 2022; Ságvári et al., 2021), the response rate was, in fact, significantly lower among the group invited to participate in the app for one day compared to those invited for seven days (11 vs. 14 %). As expected, participants invited to track their travel behavior for one day completed the study more often (71 %) than those invited to track for seven days (58 %). Respondents invited to participate for one day often extend their participation, implying that the cumulative burden of participating for multiple days is not so high. Especially if this cumulative burden is compared with the high initial burden of downloading and registering a study app.

We do find significant effects of socio-demographic variables on participation in the app and web diary. In the web diary, we find people living in highly urban areas, and people in the second income quartile (compared to the first income quartile) less likely to participate. In the app diary we find self-employed people and people not born in the Netherlands less likely to participate. People in the highest income quartile are more likely to participate in the app. Both in the app and in the web diary two-person households are more likely to participate than one-person households. Additionally, we find that one-person households were less likely to participate in the seven-day diary than in the one-day diary than persons with other household compositions. Except for the household type, we did not find any other significant relations between socio-demographic variables and the study duration.

There were only small differences in the measurement quality between the one-day and the seven-day group on the first full day. Both groups provide on average 16 h with geolocation measurements. The one-day group is slightly more actively engaged than the seven-day group on their first full day, demonstrated by a higher labeling rate (78 vs. 73 %). In the seven-day group we do see some sort

of fatigue effect, as the average number of measurements collected during the first six days is lower than that of the first full day. We do not see a strong difference in active engagement over time.

The dropout from the app study was, however, considerable, with 29 % of the one-day group discontinuing participation within 24 h and 42 % of the seven-day group dropping out before the seventh day. It is difficult to disentangle the potential reasons for dropout such as app usability, technical issues with devices, and respondent motivation. The results of the logistic regression model predicting completion imply that the app functionalities and device types did affect attrition. While not everyone uses the editing functionalities, participants with limited editing functionalities dropped out more often than those with extensive editing functionalities. We assume smartphone users are used to having extensive functionalities in apps and we received anecdotal in-app respondent feedback that not being able to edit mistakes in the diary is frustrating. Participants with a smartphone other than an iPhone or a Samsung completed the study less often. We saw that these devices had more technical issues which can add to cumulative burden. Overall, the combined response rate was 19 % (12 % for the app and 7% for the web diary) which was lower than we expected based on the previous travel app in 2018 (27 % for seven days, see Lugtig et al., 2022; McCool et al., 2021). Besides the overall declining response rates, and the recruitment strategy with a low conditional incentive, the experimental study design likely has impacted the response. The mix of modes, resulting in lengthy and complex invitation letters may have confused or demotivated respondents. Another limitation of our study was the measurement quality of the location data. Also when omitting some problematic brands and models, the amount of missing data and the drop-out were larger than in 2018.

The higher response rate for the seven day groups as opposed to the one day group is intriguing. We see two likely explanations for the lower response rate for the one-day study. The assignment of a fixed starting day hinders people from starting the study immediately after which they might simply forget to participate. Another potential explanation is that the initial burden of downloading the app, setting it up and adhering to the study protocol is too high for a one-day study so that participants might think the effort is simply not worth their time.

Another reason to argue for a longer study duration is that people have trouble following instructions regarding when to begin and finish the one-day diary. Since participants can only start to provide data upon app installation, they can provide fewer hours on the starting day. Therefore, to ensure a full day of data collection, we instructed the one-day group to begin using the app on the day following their app registration. However, some people only used the app on the day of the app installment. Moreover, not everyone invited to participate for one day used the diary app on their assigned day.

As a byproduct of our analyses, we learned that active app respondents, regardless of whether they completed the study of the duration assigned to them or not, have varying response patterns. Some did not provide any geolocation measurements after one hour, while others still did for a couple of days. Moreover, the variation in the number of geolocation measurements provided is high. Due to this variation in measurement quality among participants a different definition of study completion that takes into account the quality of the collected data, might be useful. Therefore, further research into thresholds of a sufficient number of hours and days with geolocation data to calculate valid travel statistics is necessary. This is also useful from the perspective of data minimization. Researchers and national statistical institutes want to collect minimal data (for example, to protect participants' privacy) while being able to calculate meaningful statistics. Further research is necessary in order to recommend for how many days researchers need to collect app travel diary data.

We see a few more avenues for future research and explorations. An important open question is how to combine different modes, if at all. A recommendation is to go for a harmonization of web questionnaire and app design parameters. Furthermore, future studies on the feasibility of app data collection for official statistics might benefit from a more explicit focus on burden by asking the respondents in-situ. We do have measures on how much time people spend in the app, but we did not ask about the experienced burden in the app. Furthermore, as discussed by Read (2019), it would be useful to compare the burden between mobile app data collection and traditional survey methods.

We conclude with a few recommendations based on our experiences with conducting the app-based diary study. Firstly, we advise to offer an alternative mode next to the app mode as not everyone is able or willing to use an app. In our study adding a web mode did not lower the app uptake. Additionally, the composition of the web and app respondents does not differ greatly in terms of the available socio-demographic background characteristics. Why some people prefer the app or web mode for a diary study, needs further investigation with different sets of variables. Secondly, we advise to conduct the study for one week as the willingness to participate multiple days is high, and the beforementioned advantages of collecting multiple diary days per person are compelling. Thirdly, we advise to offer extensive (editing) functionalities in the app as this does not adversely affect participant dropout rates but does have the advantage of improved data quality.

Acknowledgements Data collection was funded by the department of infrastructure from the Ministry of Transport, Public Works and Water Management, and Statistics Netherlands. We also want to thank our colleagues at CBS involved with the travel app project, including Jelmer de Groot, Mike Vollebregt, Anne Elevelt, Jonas Klingwort, and Yvonne Gootzen for help with preparing the data.

References

Allström, A., Kristoffersson, I., & Susilo, Y. (2017). Smartphone based travel diary collection: experiences from a field trial in stockholm. *Transportation Research Procedia*, *26*, 32–38. https://doi.org/10.1016/j.trpro.2017.07.006.

American Association for Public Opinion Research (2023). Standard definitions: final dispositions of case codes and outcome rates for surveys, 10th edition. https://aapor.org/wp-content/uploads/2024/03/Standards-Definitions-10th-edition.pdf

Antoun, C., & Wenz, A. (2021). *Participation metrics for accelerometer-based research*. MASS Workshop, 22.04.

Berg, N., Seferi, G., Holmøy, A., Egge-Hoveid, K., & Lund, K.-A. (2022). *The impact of a smart survey approach on participation and data quality*. Nordic Statistical Meeting, 27.02.

te Braak, P., Minnen, J., & Glorieux, I. (2020). The representativeness of Online time use surveys. Effects of individual time use patterns and survey design on the timing of survey dropout. *Journal of Official Statistics*, *36*(4), 887–906. https://doi.org/10.2478/jos-2020-0042.

Bradburn, N.M. (1978). Respondent burden. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 35–40).

Brown, L.D., Cai, T.T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*. https://doi.org/10.1214/ss/1009213286.

CBS (2023). Onderweg in Nederland (ODiN) 2022—Onderzoeksbeschrijving. https://www.cbs.nl/nl-nl/longread/rapportages/2023/onderweg-in-nederland–odin—2022-onderzoeksbeschrijving

Cernat, A., Sakshaug, J., Christmann, P., & Gummer, T. (2022). The impact of survey mode design and questionnaire length on measurement quality. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221140139.

Chatzitheochari, S., & Mylona, E. (2021). Data quality in web and app diaries: a person-level comparison. *Journal of Time Use Research*. https://doi.org/10.32797/jtur-2021-2.

Cottrill, C.D., Pereira, F.C., Zhao, F., Dias, I.F., Lim, H.B., Ben-Akiva, M.E., & Zegras, P.C. (2013). Future mobility survey: experience in developing a smartphone-based travel survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, *2354*(1), 59–67. https://doi.org/10.3141/2354-07.

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513–539. https://doi.org/10.1093/jssam/smz008.

Dillman, D.A., Sinclair, M.D., & Clark, J.R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *17*.

Earp, M., Kaplan, R., & Toth, D. (2022). Modeling the relationship between proxy measures of respondent burden and survey response rates in a household panel survey. *Journal of Official Statistics*, *38*(4), 1145–1175. https://doi.org/10.2478/jos-2022-0049.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

European Commission. Statistical Office of the European Union (2020). *Harmonised European Time Use Surveys: 2018 guidelines : re edition, 2020 edition*. Publications Office. https://doi.org/10.2785/160444.

Eurostat (2020). Household budget survey 2015 wave EU quality report. https://ec.europa.eu/eurostat/documents/54431/1966394/HBS_EU_QualityReport_2015.pdf/72d7e310-c415-7806-93cc-e3bc7a49b596 (Created January).

Faghih Imani, A., Harding, C., Srikukenthiran, S., Miller, E.J., & Nurul Habib, K. (2020). Lessons from a large-scale experiment on the use of smartphone Apps to collect travel diary data: the "city logger" for the greater golden horseshoe area. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*(7), 299–311. https://doi.org/10.1177/0361198120921860.

Fischer, F., & Kleen, S. (2021). Possibilities, problems, and perspectives of data collection by mobile Apps in longitudinal epidemiological studies: scoping review. *Journal of Medical Internet Research*, *23*(1), e17691. https://doi.org/10.2196/17691.

Flake, L., Lee, M., Hathaway, K., & Greene, E. (2017). Use of Smartphone panels for viable and cost-effective GPS data collection for small and medium planning agencies. *Transportation Research Record: Journal of the Transportation Research Board*, *2643*(1), 160–165. https://doi.org/10.3141/2643-17.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response

quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031.

Geurs, K.T., Thomas, T., Bijlsma, M., & Douhou, S. (2015b). Automatic trip and mode detection with move smarter: first results from the Dutch mobile mobility panel. *Transportation Research Procedia*, *11*, 247–262. https://doi.org/10.1016/j.trpro.2015.12.022.

Gillis, D., Lopez, A.J., & Gautama, S. (2023). An evaluation of Smartphone tracking for travel behavior studies. *ISPRS International Journal of Geo-Information*, *12*(8), 335. https://doi.org/10.3390/ijgi12080335.

Glorieux, I., & Minnen, J. (2009). How many days? A comparison of the quality of time-use data from 2-day and 7-day diaries. *Electronic International Journal of Time Use Research*, *6*(2), 314–327. https://doi.org/10.13085/eIJTUR.6.2.314-327.

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., & Crane, M. (2015). A web-based diary and companion Smartphone app for travel/activity surveys. *Transportation Research Procedia*, *11*, 297–310. https://doi.org/10.1016/j.trpro.2015.12.026.

Groves, R.M., Singer, E., & Corning, A. (2000). Leverage-Saliency theory of survey participation description and an illustration. *Public Opinion Quarterly*, *11*.

Groves, R.M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *31*.

Harding, C., Faghih, I.A., Srikukenthiran, S., Miller, E.J., & Habib, N.K. (2021). Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. *Transportation*, *48*(5), 2433–2460. https://doi.org/10.1007/s11116-020-10135-7.

Hargittai, E. (2002). Second-level digital divide: differences in people's online skills. *First Monday*, (4). https://doi.org/10.5210/fm.v7i4.942.

Hoerger, M. (2010). Participant dropout as a function of survey length in Internet-mediated university studies: implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking*, *13*(6), 697–700. https://doi.org/10.1089/cyber.2009.0445.

Hong, S., Zhao, F., Livshits, V., Gershenfeld, S., Santos, J., & Ben-Akiva, M. (2021). Insights on data quality from a large-scale application of smartphone-based travel survey technology in the Phoenix metropolitan area, Arizona, USA. *Transportation Research Part A: Policy and Practice*, *154*, 413–429. https://doi.org/10.1016/j.tra.2021.10.002.

Hu, M., Gremel, G.W., Kirlin, J.A., & West, B.T. (2017). Nonresponse and Underreporting Errors Increase over the Data Collection Week Based on Paradata from the National Household Food Acquisition and Purchase Survey. *The Journal of Nutrition*, *147*(5), 964–975. https://doi.org/10.3945/jn.116.240697.

Jäckle, A., Couper, M.P., Burton, J., & Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Survey Research Methods*.

Jäckle, A., Burton, J., Couper, M.P., & Perelli, B. (2023). *Participation of household panel members in daily burst measurement using a mobile app: effects of position of the invitation, bonus incentives, and number of daily questions*

Janssens, K.A.M., Bos, E.H., Rosmalen, J.G.M., Wichers, M.C., & Riese, H. (2018). A qualitative approach to guide choices for designing a diary study. *BMC Medical Research Methodology*, *18*(1), 140. https://doi.org/10.1186/s12874-018-0579-6.

Jonker, N., & Kosse, A. (2013). Estimating cash usage: the impact of survey design on research outcomes. *Economist*, *161*(1), 19–44. https://doi.org/10.1007/s10645-012-9200-2.

Keusch, F., & Conrad, F.G. (2022). Using Smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and Methodology*, *10*(4), 863–885. https://doi.org/10.1093/jssam/smab035.

Keusch, F., Struminskaya, B., Antoun, C., Couper, M.P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public Opinion Quarterly*, *83*(S1), 210–235. https://doi.org/10.1093/poq/nfz007.

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2020). Coverage error in data collection combining mobile surveys with passive measurement using Apps: data from a German national survey. *Sociological Methods & Research*. https://doi.org/10.1177/0049124120914924.

Keusch, F., Struminskaya, B., Kreuter, F., & Weichbold, M. (2020). Combining active and passive mobile data collection: a survey of concerns. In *Big Data Meets Survey Science*.

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., Trappmann, M., & Eckman, S. (2022). Non-participation in Smartphone data collection using research Apps. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(Supplement_2), S225–S245. https://doi.org/10.1111/rssa.12827.

Klingwort, J., Gootzen, Y., Remmerswaal, D., & Schouten, B. (2025). Algorithms versus survey response: comparing a smart survey travel and mobility app with a web diary. Manuscript submitted for publication.

Kreuter, F. (2013). Improving surveys with Paradata: introduction. In F. Kreuter (Ed.), *Improving surveys with Paradata* (1st edn., pp. 1–9). Wiley. https://doi.org/10.1002/9781118596869.ch1.

Lugtig, P. (2014). Panel attrition: separating stayers, fast Attriters, gradual Attriters, and lurkers. *Sociological Methods & Research*, *43*(4), 699–723. https://doi.org/10.1177/0049124113520305.

Lugtig, P., Roth, K., & Schouten, B. (2022). Nonresponse analysis in a longitudinal smartphone-based travel study. *Survey Research Methods*. https://doi.org/10.18148/SRM/2022.V16I1.7835.

Luiten, A., Hox, J., & de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: results of an international study across countries and surveys. *Journal of Official Statistics*, *36*(3), 469–487. https://doi.org/10.2478/jos-2020-0025.

Lynch, J., Dumont, J., Greene, E., & Ehrlich, J. (2019). Use of a Smartphone GPS application for recurrent travel behavior data collection. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(7), 89–98. https://doi.org/10.1177/0361198119848708.

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An app-assisted travel survey in official statistics: possibilities and challenges. *Journal of Official Statistics*, *37*(1), 149–170. https://doi.org/10.2478/jos-2021-0007.

Molloy, J., Castro, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B., & Axhausen, K.W. (2022). The MOBIS dataset: A large GPS dataset of mobility behaviour in Switzerland. *Transportation*. https://doi.org/10.1007/s11116-022-10299-4.

Nahmias-Biran, B., Han, Y., Bekhor, S., Zhao, F., Zegras, C., & Ben-Akiva, M. (2018). Enriching activity-based models using Smartphone-based travel surveys. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(42), 280–291. https://doi.org/10.1177/0361198118798475.

Ochoa, C. (2022). Willingness to participate in geolocation-based research. *PLOS ONE*, *17*(12), e278416. https://doi.org/10.1371/journal.pone.0278416.

Patterson, Z., & Fitzsimmons, K. (2016). Datamobile: Smartphone travel survey experiment. *Transportation Research Record: Journal of the Transportation Research Board*, *2594*(1), 35–43. https://doi.org/10.3141/2594-07.

Peytchev, A. (2009). Survey Breakoff. *Public Opinion Quarterly*, *73*(1), 74–97. https://doi.org/10.1093/poq/nfp014.

Peytchev, A., & Peytcheva, E. (2017). Reduction of measurement error due to survey length: evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*, 361–368. https://doi.org/10.18148/SRM/2017.V11I4.7145.

Peytcheva, E., & Groves, R.M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, *10.*, .

Prelipcean, A.C., Susilo, Y.O., & Gidófalvi, G. (2018). Collecting travel diaries: current state of the art, best practices, and future research directions. *Transportation Research Procedia*, *32*, 155–166. https://doi.org/10.1016/j.trpro.2018.10.029.

Publications Office (2020). *Harmonised European time use surveys: 2018 guidelines : re edition, 2020 edition*. https://doi.org/10.2785/160444.

R Core Team (2021). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Read, B. (2019). Respondent burden in a mobile App: evidence from a shopping receipt scanning study. *Survey Research Methods.*, , .

Revilla, M. (2022). How to enhance web survey data using metered, geolocation, visual and voice data? *Survey Research Methods*. https://doi.org/10.18148/SRM/2022.V16I1.8013.

Revilla, M., & Höhne, J.K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538–545. https://doi.org/10.1177/1470785320943049.

Revilla, M., & Ochoa, C. (2017). Ideal and Maximum Length for a Web Survey. *International Journal of Market Research*, *59*(5)., .

Revilla, M., Couper, M.P., & Ochoa, C. (2019). Willingness of online panelists to perform additional tasks. *Methods, data*. https://doi.org/10.12758/MDA.2018.01.

Roddis, S., Winter, S., Zhao, F., & Kutadinata, R. (2019). Respondent preferences in travel survey design: an initial comparison of narrative, structured and technology-based travel survey instruments. *Travel Behaviour and Society*, *16*, 1–12. https://doi.org/10.1016/j.tbs.2019.03.003.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003.

Safi, H., Assemi, B., Mesbah, M., Ferreira, L., & Hickman, M. (2015). Design and implementation of a Smartphone-based travel survey. *Transportation*

Research Record: Journal of the Transportation Research Board, *2526*(1), 99–107. https://doi.org/10.3141/2526-11.

Ságvári, B., Gulyás, A., & Koltai, J. (2021). Attitudes towards participation in a passive data collection experiment. *Sensors*, *21*(18), 6085. https://doi.org/10.3390/s21186085.

Scherpenzeel, A. (2017). Mixing Online Panel Data Collection with Innovative Methods. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 27–49). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-15834-7_2.

Schlich, R., & Axhausen, K.W. (2003). Habitual travel behaviour: evidence from a six-week travel diary. *Transportation.*, , .

Schmidt, T. (2014). Consumers' recording behaviour in payment diaries—empirical evidence from Germany. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2014-00008.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *12*, 14.

Schouten, B., Remmerswaal, D., Elevelt, A., de Groot, J., Klingwort, J., Schijvenaars, T., Schulte, M., & Vollebregt, M. (2024). *A smart travel survey results of a push-to- smart field experiment in the Netherlands*. Discussion Paper, Statistics Netherlands.

Shankari, K., Bouzaghrane, M.A., Maurer, S.M., Waddell, P., Culler, D.E., & Katz, R.H. (2018). e-mission: an open-source, Smartphone platform for collecting human travel data. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(42), 1–12. https://doi.org/10.1177/0361198118770167.

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, *142*(1), 201–211. https://doi.org/10.1016/j.jspi.2011.07.008.

Silberstein, A.R., & Scott, S. (1991). Expenditure diary surveys and their associated errors. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.), *Wiley series in probability and statistics* (pp. 303–326). Wiley. https://doi.org/10.1002/9781118150382.ch16.

Stanley, K., Yoo, E.-H., Paul, T., & Bell, S. (2018). How many days are enough?: capturing routine human mobility. *International Journal of Geographical Information Science*, *32*(7), 1485–1504. https://doi.org/10.1080/13658816.2018.1434888.

Storesund Hesjevoll, I., Fyhri, A., & Ciccone, A. (2021). App-based automatic collection of travel behaviour: a field study comparison with self-reported behaviour. *Transportation Research Interdisciplinary Perspectives*, *12*, 100501. https://doi.org/10.1016/j.trip.2021.100501.

Struminskaya, B., & Keusch, F. (2020). Editorial: from web surveys to mobile web to apps, sensors, and digital traces. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2020-00015.

Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J.K. (2020). Augmenting surveys with data from sensors and Apps: opportunities and challenges. *Social Science Computer Review*. https://doi.org/10.1177/0894439320979951.

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2021). Understanding willingness to share Smartphone-sensor data. *Public Opinion Quarterly*, *84*(3), 725–759. https://doi.org/10.1093/poq/nfaa044.

Svaboe, G.B.A., Tørset, T., & Lohne, J. (2021). Recruitment strategies in app-based travel surveys: methodological explorations. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3988501.

Tienda, M., & Koffman, D. (2021). Using paradata to evaluate youth participation in a digital diary study. *Social Science Computer Review*, *39*(4), 666–686. https://doi.org/10.1177/0894439320929272.

Toepoel, V., & Lugtig, P. (2022). Modularization in an era of mobile web: investigating the effects of cutting a survey into smaller pieces on data quality. *Social Science Computer Review*, *40*(1), 150–164. https://doi.org/10.1177/0894439318784882.

Toepoel, V., Lugtig, P., & Schouten, B. (2020). Active and passive measurement in mobile surveys. *The Survey Statistician*, *82*, 14.

Vercruyssen, A., Roose, H., Carton, A., & Putte, B.V.D. (2014). The effect of busyness on survey participation: being too busy or feeling too busy to cooperate? *International Journal of Social Research Methodology*, *17*(4), 357–371. https://doi.org/10.1080/13645579.2013.799255.

Wenz, A. & Keusch, F. (2022). The second-level smartphone divide: A typology of smartphone use based on frequency of use, skills, and types of activities. *Mobile Media & Communication*, 1–25. https://doi.org/10.1177/20501579221140761.

Wenz, A., Jäckle, A., & Couper, M.P. (2019). Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods*, *22.*, .

Winkler, C., Meister, A., & Axhausen, K.W. (2024). The TimeUse+ data set: 4 weeks of time use and expenditure data based on GPS tracks. *Transportation*. https://doi.org/10.1007/s11116-024-10517-1.

Yan, T., & Williams, D. (2022). Response burden—review and conceptual framework. *Journal of Official Statistics*, *38*(4), 939–961. https://doi.org/10.2478/jos-2022-0041.

Yazdizadeh, A., Patterson, Z., & Farooq, B. (2019). An automated approach from GPS traces to complete trip information. *International Journal of Transportation Science and Technology*, *8*(1), 82–100. https://doi.org/10.1016/j.ijtst.2018.08.003.

Zhao, F., Pereira, F.C., Ball, R., Kim, Y., Han, Y., Zegras, C., & Ben-Akiva, M. (2015). Exploratory analysis of a Smartphone-based travel survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, *2494*(1), 45–56. https://doi.org/10.3141/2494-06.