Anne Elevelt

Smart(phone) Surveys

Smart(phone) Surveys

Smart(phone) Surveys

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht op gezag van de rector magnificus,
prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van
het college voor promoties in het openbaar te verdedigen
op vrijdag 16 april 2021 des middags te 12.45 uur

door

Anne Elevelt

geboren op 5 juni 1993
te Eindhoven

Promotor:	Prof. dr. P.G.M. van der Heijden
Copromotoren:	Dr. P. Lugtig
	Dr. V. Toepoel

Beoordelingscommissie: Prof. dr. J.W.M. Das
Prof. dr. E.D. De Leeuw
Prof. dr. F. Keusch
Prof. dr. J.G. Schouten
Prof. dr. T. Van Der Lippe

Smart(phone) Surveys

Proefschrift Universiteit Utrecht, Utrecht

Author: Anne Elevelt
Illustrations: Olly Kava
Graphic Design: Nino Schöningh
Printing: Ridderprint
ISBN: 978-94-6416-400-8

Table of Contents

Chapter 1

Introduction

Almost everyone has a smartphone. In the Netherlands, 92.1% of the total population owned a smartphone with Internet access in 2019, and for individuals between 18 and 65 years this percentage was 95.2% (Statline, 2019). Of course, the Internet and smartphone penetration rates differ per country but worldwide the share of internet traffic that can be assigned to smartphones is already larger than the share of internet traffic that can be assigned to computers (Clement, 2020; Kantar Public Brussels, 2019).

People use their smartphone for much more than just calling and carry their smartphones everywhere they go. I think it's safe to say that your smartphone lies somewhere within five meters of you while you are reading this. And that you'll unlock your phone before you have reached the end of this chapter[1].

Your smartphone probably knows more about certain aspects of your life than you do yourself, or at least it knows them more precisely. It knows for example: how far you've travelled yesterday; how many minutes you spent on Instagram last week; how often you played your favorite song (or "*What is Love*") last year; your friends' birthdays and next week's appointments; how many steps you take on average per day; how physically active you are. This made me and many other methodologists think: why wouldn't you use smartphones for social science research?

[1] Which may be due to the boring nature of this dissertation. Perks of being a methodologist.

1.1 From Online Surveys to Mobile Web Surveys

With the popularization of the Internet, online surveys have become the dominant mode of survey data collection because they are time and cost efficient. Online surveys are relatively cheap because they do not require interviewers, printing, postage or transcription. Online surveys are quickly conducted because large numbers of respondents can answer survey questions in a short amount of time. Traditionally, (almost) all online surveys were completed on PCs or laptop computers.

This is shifting; Nowadays all online surveys are mixed-device surveys (Beuthner, Daikeler, & Silber, 2019; Couper, Antoun, & Mavletova, 2017; De Leeuw & Toepoel, 2018; Toepoel & Lugtig, 2015). In (most) online surveys, respondents are free to choose which device they want to use to answer a survey. More and more people are completing these surveys on their smartphone. The usage of smartphones and various other devices among survey respondents leads to a variety of screen sizes, data entry methods (keyboard or touch screen), operating systems, and web browsers that influence how a web survey questionnaire is displayed on the devices and how respondents complete the questionnaire (Antoun, Katz, Argueta, & Wang, 2017).

This is a major challenge for survey designers, as the goal is to enable all respondents, no matter on which device they answer, to achieve a pleasant and preferably similar survey experience. Survey designers need to rethink the design conventions for online surveys that are traditionally based on PC's (Antoun et al., 2017; Antoun & Cernat, 2019). Smartphones pose a greater challenge to design than computers, as smartphones have smaller screens and navigate through fingers/touch instead of a more precise mouse. Mobile-first design results in a mobile-optimized design, which seems like a logical approach since screen size is one of the most important issues related to usability. If a survey looks good on a smartphone, it should look good on all other devices as well.

1.2 Going Beyond the Traditional Survey (with Smartphones)

Smartphones also bring opportunities to go beyond the traditional survey and collect new types of data that cannot be collected via computers. We see three main opportunities here to enhance or extend measurement; the promise of "anytime, anywhere" measurement, the use of research apps and the collection of sensor data.

First, smartphones enable respondents to access the Internet "anytime, anywhere" and thus allow them to answer questionnaires independent of time and location, e.g. in a train or at the airport (Couper, Antoun, & Mavletova, 2017). This offers possibilities for measurement through Experience Sampling Methods, where respondents are repeatedly

sent triggers to report on their subjective experiences (e.g. feelings, behavior) in that moment (Harari et al. 2016). Or location-based measurement, in which respondents can be prompted to provide survey responses at fixed locations, such as festivals or stores. Second, smartphones enable researchers to conduct their surveys through research apps. Research apps are pieces of software that can be used for data entry and sensor data collection. Apps run in the background and can thus easily collect and store data over a longer period of time, making them particularly interesting for surveys with longer data collection periods. In addition, reminders can easily be sent through an app as well as pop-up questions that need to be answered on the spot.

Third, smartphones incorporate a large number of sensors (e.g. accelerometers, GPS, light and proximity sensors) which can be logged passively, providing a large and detailed set of measurements about respondents and their environment. All these sensors can be used for research purposes. This enables researchers to collect high-intensity data passively, that is, there is no respondent activity needed after giving a one-time permission to share sensor measures with the researchers. Additional data can also be collected actively, for example when respondents are asked to take pictures or scan barcodes.

The opportunities mentioned, can simplify the response task for respondents. Sensor data can (partly) replace survey questions. For example, in widely used diary surveys respondents are asked to for example report on their time use, mobility behavior, consumption behavior or food consumption for several days. These diaries are burdensome to complete and respondents may be less willing to put effort in the diary over time, resulting in higher nonresponse and lower accuracy over days (Chatzitheochari et al., 2017). The diary completion process can be eased by introducing data entry through the scanning of receipts in a Household Budget Survey using the smartphone's camera (Jäckle et al., 2019; Wenz et al., 2019; Statistics Netherlands, 2019). Or by detecting trips via GPS sensors in a mobility survey (McCool et al., *in press*). Another advantage is that sensors potentially generate better data than respondents can provide themselves. This provides research with substantive information which would be either impossible or inaccurate to collect from self-reports alone. For example, health surveys suffer from inaccurate statistics of how physically active people are. Respondents find it difficult to remember how much time they are physically active, let alone how intense these periods of activity are. Smartphone surveys or wearables can provide more accurate statistics on this.

1.3 Methodological Questions

This all seems very promising, but many methodological questions arise about how to conduct such smartphone surveys.

First, methodological questions related to representation of the population arise. Participants may not be willing to share sensor data or only a specific group may be willing to share sensor data, inducing bias. Besides, not everyone owns a smartphone, and not all smartphone owners are able to use apps. Some studies loan respondents a mobile device to overcome these coverage problems, but this approach is not common (Scherpenzeel, 2017). There are also technical reasons for nonresponse related to the willingness to download and use smartphone apps, such as battery life and availability of storage space on the device. Furthermore, participants may forget to recharge their smartphones, or to carry them, interrupting data collection. Lastly, participants may use several phones per day, for example a corporate-leased work phone and a private phone, which may restrict data collection to work or out-of-work times.

Second, methodological questions related to measurement arise. We do not know how useful the additional data collected via sensors and apps are. Smartphone sensors are not designed for collecting data for research, but for common phone functions, and have therefore limited accuracy. Geographical location data can for example contain measurement error. GPS coverage varies and GPS accuracy depends on satellite geometry, signal blockage, atmospheric conditions, and receiver design features or quality (in other words: the phone itself). Geographical location data may be inaccurate as GPS trackers may lose signal indoors, in trains or in areas with many tall buildings leading to positioning errors.

Third, technical and ethical questions arise. Because of the large quantity and privacy sensitive nature of data that can be collected, there is a need for procedures to collect, save, manage, manipulate and analyze the data. Ethical challenges also cover questions on how to obtain truly informed consent and protecting participant privacy and anonymity.

1.4 The Total Survey Error Framework

The problems with representation and measurement in smartphone surveys can be better understood when they are embedded in the Total Survey Error Framework: It is important to understand this framework to understand the outline and further chapters. The TSE framework is the dominant paradigm used by survey methodologist to describe all error sources that can affect the estimates arising from the survey. These errors arise in all phases of the survey lifecycle: design, implementation (collection & processing), and data analysis. The goal of survey methodologists is optimizing surveys by minimizing the errors.

Groves et al. (2004) combined aspects of the TSE framework in a flowchart as shown in Figure 1.1. They distinguish between errors related to the representation of the population, and errors related to the measurement of the survey construct.

The representation side of the framework focuses on population statistics, not on individual responses. Errors related to representation are coverage error, sampling error, and nonresponse error. First, coverage errors arise when a gap exists between the target population and the sampling frame. Sometimes the target population does not have a sampling frame that matches it perfectly. For example, in a sampling frames based on mobile phone numbers there may be undercoverage of older participants as they are less likely to own a smartphone.

Second, sampling errors are inevitable in every sample surveys as differences between the sampling frame and the sample always occur. It is simply infeasible to measure all individuals in the sampling frame. Therefore, a sample of persons is selected, introducing deviations from the sampling frame. When the sample drawn from the sampling frame is too small though, the target population is not represented very well, resulting in a large sampling variance. Furthermore, systematic flaws (or biases) in the sampling method can increase sampling errors when some members of the sampling frame are given no chance (or reduced chance) of selection.

Third, nonresponse errors arise when a gap exist between the sample and the respondent pool. Nonresponse can be described as the failure to obtain measurements on sampled units. Nonresponse occurs at the unit level when sample members cannot be contacted or refuse cooperation with the survey request or additional data collection. Nonresponse at the item level when respondents do not provide an answer to one or multiple questions. Non-consent is a specific type of nonresponse in smartphone studies where respondents refuse to share additional data. Nonresponse in itself is not a big problem as long as respondents and nonrespondents do not systematically differ on the variables of interest. Nonresponse error arise when the values of the statistics computed based only on respondents data differs from those of the entire sample data.

Errors related to measurement are specification errors, measurement errors and processing errors. First, specification errors are related to construct validity; the extent to which the measure is related to the underlying construct. Construct validity refers to the extent to which a survey actually measures the construct of interest. For example, to what extent is an IQ questionnaire actually measuring intelligence?

Second, measurement errors are the differences between the true value and the value provided by the respondent. Measurement errors are influenced by many factors but mostly by mode, questionnaire design and respondent behavior. Almost all survey design choices impact measurement error, for example question wording, scales and visual presentation influence response accuracy. Non-optimized smartphone surveys may for example induce measurement error because not all answer options fit on the screen

appropriately. Respondent behavior can also cause error because of the way respondents cognitively process and answer questions.

Third, processing errors occur when the variable used in the estimation is different than that provided by the respondents. Processing errors can occur after the data have been collected. For example, sometimes it is desired to recode open questions into numbered classes. Different persons coding these text answers can make different judgments about how to classify the text answers causing unwanted variability in the results.

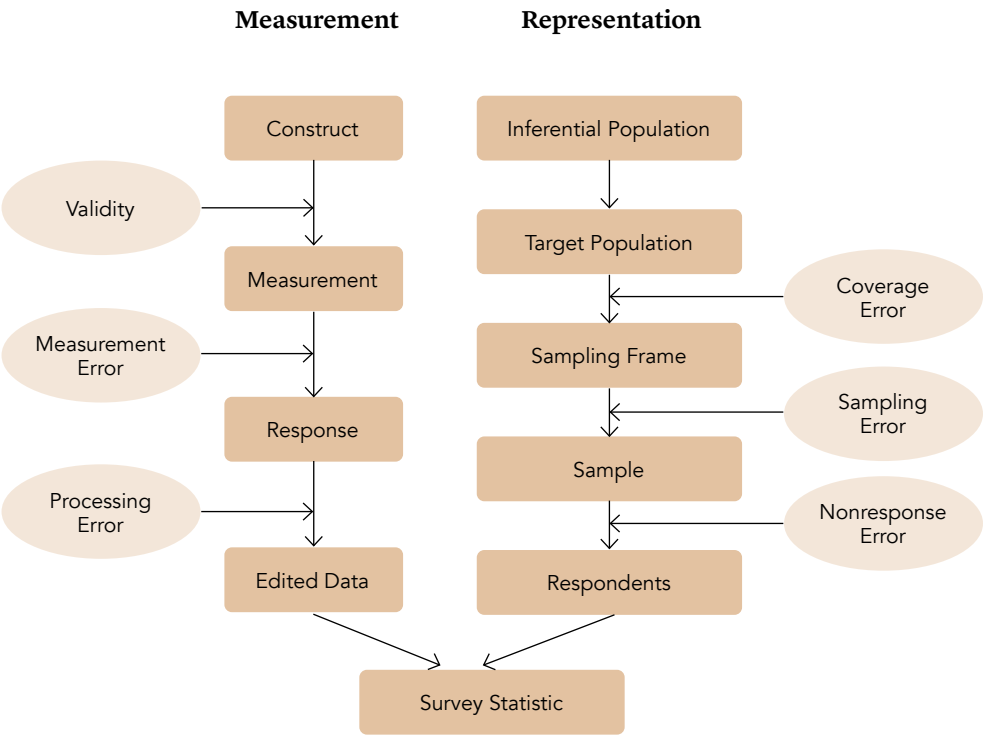


Figure 1.1 Total Survey Error Framework (Groves et al., 2004; Groves & Lyberg, 2010).

1.5 The Total Survey Error Framework and Smartphones

The steps in the TSE framework become more complicated with the introduction of smartphones and new additional data types. The TSE framework has been adapted over time to include these new methods of data collection (Amaya, Biemer & Kinyon, 2020; Groves & Lyberg, 2010); when supplementing survey data with others sources,

understanding each source's main errors allows researchers to take advantage of the strengths of each data source and create an optimal, combined end product (Biemer, 2016).

Smartphone surveys face a trade-off between measurement quality and response rates; using smartphones probably decreases errors on the measurement side of the TSE but increases errors on the representation side of the TSE. The question is what the total effect of smartphone surveys is on the TSE; whether the larger error in the representation part outweighs the smaller error in measurement.

Let us first take a look at the representation side. Nonresponse errors are a large concern as smartphone surveys involve more steps than traditional online surveys. Respondents may need to download an app, give consent to share sensor data, perform additional tasks, and answer pop-up questions. In all these steps, respondents may drop out. Nonresponse bias could be introduced or accumulate at every step. There are many more types and gradations of nonresponse and nonresponse bias.

Now, let us take a look at the measurement side. As described earlier, smartphone surveys and sensor data are expected to increase measurement quality. A key challenge for measurement is that the survey data and sensor data are designed for different purposes though. Researchers need to critically rethink their operationalizations; using one data source (i.e. sensor data) to replace another (i.e. survey questions) requires knowledge of what these data sources actually measure and how they measure it. Specification errors may occur when researchers know too little about what these data sources actually measure. Processing or coding errors may occur when the data sources are combined or aligned incorrectly.

1.6 Outline

In this dissertation we investigate the effect of smartphone surveys in terms of reducing or enlarging TSE components. The remainder of this dissertation follows the TSE survey lifecycle; We start with consent in Chapter 2, then nonresponse (bias) in Chapter 3, and measurement error in Chapter 4. We end with an experimental application of a smartphone study in Chapter 5, in which we bring measurement and representation error together.

In Chapter 2 we perform a systematic review and meta-analysis to investigate how to improve the effectiveness of the consent to data linkage question. We assess experiments on different aspects of the consent question. This chapter does not only consider the consent question to sensor/passive data linkage but also to other types of data.

In Chapter 3 we study nonresponse and nonresponse bias in the smartphone-only version of the Dutch Time Use Survey (TUS). We investigate survey response rates, predictors of nonresponse and nonresponse bias in a smartphone-only TUS conducted in the Dutch probability-based LISS panel. Panel members were asked to participate in several tasks varying in intrusiveness and burden (completing surveys, a diary, sharing sensor data and answering pop-up questions). Nonresponse bias could be introduced or accumulate at every step. Subsequently we examine whether this nonresponse influences our survey estimates, resulting in nonresponse bias.

In Chapter 4 we focus on measurement error when collecting sensor data in a smartphone survey. We address the methodological challenges of analyzing and integrating GPS location data with self-reported time use diary data. We assess whether the passive recording of participants' geographical locations is sufficient to establish their functional locations (the natural functions of the locations). We propose a method for analyzing GPS data, to integrate the location and survey data, and to explain variability.

In Chapter 5 we investigate both representation and measurement in an innovative and experimental study on the use of sensor data in fitness and health research. We explore the feasibility of fitness tasks in self-administered smartphone survey using acceleration data. We investigate respondents' compliance with these tasks and whether we could validate compliance and performance on these tasks using acceleration data, measured by respondents' smartphones.

Chapter 6 summarizes and discusses the main outcomes from the studies conducted in Chapter 2 -5. Furthermore, based on the strengths and weaknesses of the studies described in this dissertation, suggestions are provided for future survey methodological research, and recommendations are given for actual survey practice. Lastly, my personal view on the future of smartphone research is presented in this chapter.

Chapter 2

Consent to Data Linkage in Surveys: A Descriptive Review and Meta-Analysis.

Elevelt, A., Toepoel, V., & Lugtig, P. (2020). *Submitted to Public Opinion Quarterly.*

Author contributions: AE, VT and PL contributed to the study concept and design. AE organized the data collection, selected the included studies, and extracted the data. VT and PL double coded a part of the literature and data-extraction. AE performed the analyses and wrote the first draft of the paper. VT and PL critically reviewed the paper.

Abstract

The widespread adoption of digital technologies is expanding opportunities for survey researchers to enhance survey research with other data sources, such as administrative records, social media data, biodata or sensor data. A crucial element enabling successful research with linked datasets, however, is to obtain respondents' consent in order to avoid significant amounts of missing data and to minimize the risk of consent bias. Unfortunately, there is still a large variability in consent rates among different studies, which cannot be fully explained and which hinders the full exploitation of data linkage potential. Our aim is to improve our understanding of the circumstances that lead to low or high consent rates.

To achieve this goal, we conducted a systematic review and meta-analysis to identify modifiable aspects of the consent to the data linkage request that influence consent rates. A systematic literature search (following the PRISMA guidelines) of six databases yielded 45 eligible manuscripts. An inventory of all conducted experiments in these manuscripts revealed a large variation in the aspect of the consent question covered. We performed a network meta-analysis for the two most-covered aspects (i.e., sponsorship and question wording). All other categories were systematically reviewed. Our results show that how researchers ask consent questions matters greatly for the achievement of high consent rates. Sponsorship, incentives, mode, position, relevancy, study duration and opt-out also affect consent rates. The results from this study can be useful in designing future consent protocols.

2.1 Introduction

The widespread adoption of digital technologies is expanding opportunities for researchers to enhance surveys with other data sources (e.g., Elevelt, Bernasco, Lugtig, de Ruiter & Toepoel; Höhne, Revilla & Schlosser, 2019; Link et al., 2014; Lugtig & Schouten, *in press;* Struminskaya et al., 2020a). For example, researchers currently use sensor data (e.g., locations, movement data), biomarker data (e.g., blood, saliva), social media data (e.g., Facebook, Twitter) or administrative records (e.g., employment data, health records, social security records) in conjunction with survey data to better understand people's attitudes and behaviour. Each data source has its own strong and weak points in regard to making inferences about a variable of interest, and combining data sources could potentially reduce errors (Biemer & Lyberg 2003; Biemer 2010).

Linking survey data to other data sources has two main advantages for researchers. First, by linking survey data to other data sources, there is more information available on the surveyed units without increasing survey burden. Several high-profile surveys, such as the UK Household Longitudinal Study (Buck & McFall, 2011) and the German Study "Labour Market and Social Security" (PASS; Trappmann, Beste, Bethmann, & Müller, 2013), already engage in linking survey data to one or more administrative data types (e.g., employment, education, health). Second, these other data sources contain important substantive information, which would be either impossible or inaccurate to collect from respondent self-reports alone. Such linked data can make survey questions obsolete because the quality of the other data sources is better. For example, saliva samples, which can be collected noninvasively by participants themselves in a self-administered survey, provide an assessment of physiological indications of stress, immune function and infectious disease, drug use and reproductive function (Hofman, 2001; Lindau & McDade, 2008). Alternatively, in the Statistics Netherlands Mobility study, passive location data collection replaces parts of a self-administered travel diary, which leads to a higher accuracy of the travel statistics (e.g., mode of transportation) (Smeets, Lugtig, & Schouten, under review).

Linking data is not always easy. To link survey data to other data sources, respondents' permission is compulsory. The General Data Protection Regulation (GDPR) regulates the protection of natural persons with regard to the collection, sharing and processing of personal data. The GDPR further states that respondents need to be informed about how their linked data will be used and how their privacy is protected. Respondents need to actively give permission to process and link their data, and this permission can be revoked by the respondent at any time.

Achieving high consent rates is a crucial step to avoid significant amounts of missing data

and minimize the risk of bias in inferences obtained from the linked data. Unfortunately, there is still a large variability in consent rates among different studies, which hinders the full exploitation of data linkage potential. This variability is remarkable and cannot yet be fully explained.

To improve our understanding of the circumstances that lead to low or high consent rates, this paper presents a systematic review and meta-analysis to identify modifiable aspects of the consent to data linkage requests that influence consent rates. Building a body of evidence for how these different aspects exactly influence consent can hopefully help to improve the effectiveness of the consent question and increase consent rates. Knowing more will allow us to ask better linkage questions, which will lead to improved linkage rates and more informative studies.

The remainder of this article is structured as follows. We start with a discussion of the consent to data linkage literature, reviewing possible causes in variation to consent rates. Then, we discuss how we conduct a literature review (i.e., eligibility criteria, study flow and study selection procedure) to systematically collect causal empirical evidence on consent rates. In the results section, we first present the results of the meta-analytical models applied to a subset of the studies identified, followed by a descriptive review of the aspects that were covered by too few studies to be included in the meta-analysis. We end with conclusions, a discussion of limitations of our study and recommendations for how to improve consent rates to data linkage.

2.2 Background

Central in the literature on consent rates is the difference between hypothetical willingness and actual consent. While hypothetical willingness refers to respondents' general disposition to share data, actual consent refers to respondents' actual sharing of the data. Many studies have investigated hypothetical willingness, while fewer have studied actual consent for linkage. Hypothetical willingness may be higher (Struminskaya, 2020a) or may be more variable across different types of data (Kreuter et al., 2018) than actual consent rates. Furthermore, people who indicate their consent to hypothetical willingness do not always provide actual consent. Scherpenzeel (2017) found, for example, that 37% of the LISS Panel respondents stated willingness to share sensor data (e.g., their geolocation and accelerometer data), whereas 81% of those actually shared these data.

Previous research has also shown large variability in consent rates across studies, populations, and tasks. For example, a systematic review on consent to administrative data linkage showed that consent rates varied between 39% (for a study on 69-year-old patients from a stroke registry) and 97% (for a study on a middle-aged general population

survey) (Da Silva et al, 2012). In a study across Netquest panelists from seven different countries, Revilla et al. (2016) found a hypothetical willingness (respondents who were *"for sure"* willing) of 17% (sharing GPS position by tablet users in Portugal) to 51% (taking pictures by smartphone users in Mexico). In a pilot study in the LISS Panel, 15 and 19% of the panel members returned biomarker data (e.g., blood and saliva samples) (Avendano, Scherpenzeel, Mackenbach, 2011). In a similar study conducted in the Relationship Dynamic and Social Life (RDSL) study, an ongoing Internet diary study of young American women, however, 65% of the respondents (150 panelists who reported the end of a romantic relationship) returned biodata (saliva samples) (Gatny, Couper, & Axinn, 2013).

In short, consent rates vary widely across studies, and it is difficult to explain or understand why these consent rates vary. The population or specific data linkage question does not seem to explain all of the variation. Anecdotal evidence suggests that the context of the request itself is a large determinant of the likelihood of consent to data linkage in a survey.

The context of the consent question includes many aspects, such as how and when the question is delivered, the topic of the survey itself, and importantly, how the wording of the consent request is framed. For example, mentioning that data linkage leads to benefits for the study or the respondent seems to increase respondents' willingness to do so (Fobia et al., 2019; Sakshaug and Kreuter, 2014; Pascale, 2011). In addition, positioning the consent question at the beginning of the survey may increase respondents' likelihood of consent (Sakshaug et al., 2013; 2015; 2019b).

How to effectively optimize asking for consent is still unclear; however, findings from research into the wording of the consent question, for example, still yield only small improvements, conflicting results or subgroup effects (Kreuter, Sakshaug, and Tourangeau, 2015; Sakshaug, Schmucker, Kreuter, Couper & Singer, 2017; Sakshaug et al., 2019). Researchers do not know how to modify the consent question for optimal consent rates. The systematic review and meta-analysis described here were designed to address the question of what works and what does not when asking for consent to data linkage.

Meta-analysis, a statistical procedure to integrate the results of previous studies, is a powerful tool to summarize the empirical knowledge in a specific field and is often used to inform policy or practice (Cehovin, Bosnjak, & Lozar Manfreda, 2019). A key benefit of meta-analysis is that the aggregation of information cancels out sampling errors of individual studies, leading to a higher level of precision, statistical power and validity (Littell, Corcoran, and Pillai 2008; Koricheva and Gurevitch 2013). This allows for the generalization of findings.

2.3 Methods

This review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009).

2.3.1 Eligibility Criteria

We prespecified our inclusion and exclusion criteria and included sources that implemented an experiment (or RCT) on the consent question for data linkage. Furthermore, we were interested in respondents, so we excluded articles about researchers or research institutes sharing research data. Finally, only articles that were written in English were included.

We used the following definitions of the inclusion criteria:

- **Experiment.** A scientific test that is done in order to discover what happens to something in particular conditions. We excluded observational studies that compared nonconsenters with consenters. We also excluded qualitative studies.
- **Consent.** Consent occurs when participants are asked to affirm that they understand the research procedure and agree voluntarily to the proposal (e.g., to share additional data).
- **Data Linkage.** Data linkage is a part of the process of data integration –linkage combines the input sources (e.g., sample surveys and administrative data). We therefore excluded studies that asked respondents' consent to participate within a randomized controlled trial (RCT) or medical experiment alone.

2.3.2 Literature Search Strategy

SCOPUS, PubMed, Web of Science, Embase, Cochrane and PsychInfo were searched from the 5th of March 2019 to the 16th of April 2019. The searches were undertaken by a research assistant using 108 combinations of keywords within the title, abstract, and full text.

The keywords were a combination of the following:

1. *experiment/clinical trial/RCT*
2. *consent/willingness/permission*
3. *"data/record + linkage/linking/joining/sharing/augmentation/blending"*

In addition, we sent a call for papers in online discussion lists relevant to survey methodologists (SRMSNet, AAPOR mailing list, ESRA mailing list) and posted a request on Twitter.

2.3.3 Study Flow

The database search in March 2019 initially yielded 2728 results, and the call for papers via the mailing lists yielded another 15 records. The identification of duplicates revealed 1035 duplicate results among the initial 2743 results. Consent to data linkage, especially passive and biodata linkage, is a relatively new and innovative topic; therefore, many projects are still in the pipeline. Therefore, we conducted a search update at the end of the first search and coding process (June 2020) to ensure that we captured all relevant and recent papers. This search update was limited to 2019 and January to June 2020 and yielded an additional 672 results. One further manuscript was sent to the authors several months after the previous call for papers and is included in the search update. The identification of duplicates from the 2020 search revealed 168 duplicate results.

After deduplication, the abstracts of the remaining 2213 manuscripts were evaluated in accordance with the inclusion criteria. If it was possible to identify a manuscript as ineligible beyond any doubt, it was excluded. In all other cases, manuscripts were scheduled for full-text evaluation.

The PRISMA flowchart (Moher et al. 2009) in Figure 2.1 illustrates the study flow and selection process.

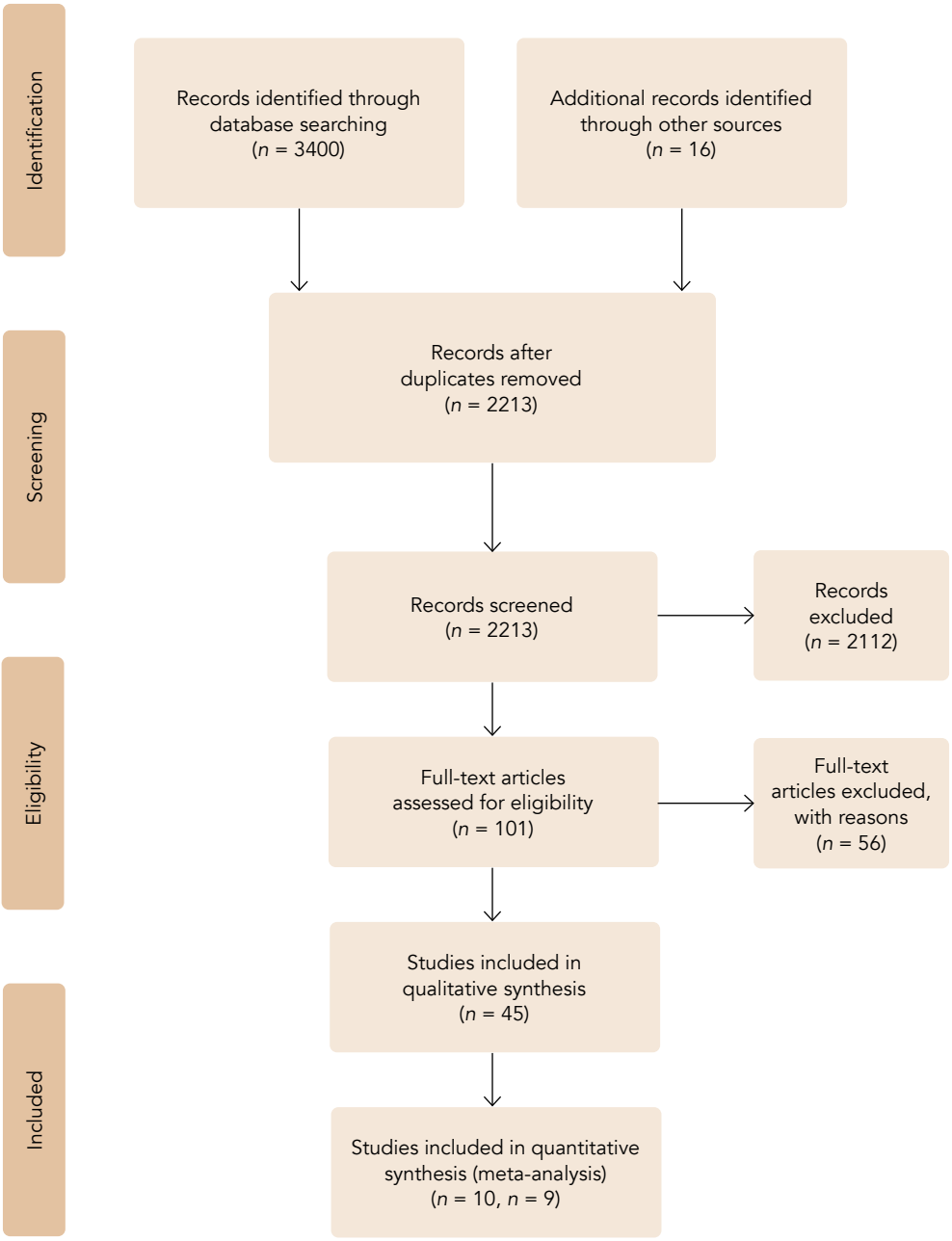


Figure 2.1 PRISMA flow diagram.

2.3.4 Study Selection

The three authors and a research assistant independently evaluated 15% of the title/abstract selection of the first search (*n* = 250) for inclusion. The interrater agreement was 92%, with Fleiss Kappa = .57, indicating moderate agreement[2]. Disagreements were discussed and resolved through consensus. Afterwards, the research assistant undertook the remainder of the title/abstract selection and discussed all abstracts with ambiguity (*n* = 51) with the first author until agreement was reached.

Subsequently, both the first author and the research assistant performed the full text selection based on the inclusion criteria *(experiment, consent,* and *data linkage)*. They coded the first 9 manuscripts together. Then, they coded the other 57 manuscripts separately, with an interrater agreement of 78.9% and Cohen's Kappa of .58, indicating moderate agreement . Disagreements (*n* =12) were discussed and resolved through consensus; 9 of these manuscripts were included.

The abstract and full text selections of the search update were performed solely by the first author. During the abstract appraisal, 2112 abstracts were excluded, and 101 manuscripts remained. The full-text content of the remaining 101 manuscripts was evaluated according to the inclusion criteria. Of the 101 manuscripts, 56 were excluded and 45 were eligible, consisting of three book sections, one presentation and 41 journal papers.

2.3.5 Meta-Analytic Procedure

We classified all experiments in the 45 eligible manuscripts. There are many aspects of the consent questions and procedure that can be varied; we summarized the experiments per type of experiment (e.g., specific framing of request, position in questionnaire). Table 2.1 presents the full inventory. To perform a network meta-analysis, we aimed to have at least ten manuscripts within every specific type of experiment. All other categories were systematically reviewed.

Network meta-analysis, in the context of a systematic review, is a meta-analysis in which multiple treatments (that is, three or more) are compared using both direct comparisons of interventions within randomized controlled trials and indirect comparisons across trials based on a common comparator (Li et al., 2011). Network meta-analysis can also take into account the effect of multiple treatment arms in one study.

[2] Fleiss Kappa takes the number of categories into account (which is only two) and is therefore "only" moderate despite the high interrater agreement.

All pairwise comparisons and risk differences are noted in a separate row in the datafile. Based on the full texts, two of the authors coded and extracted the consent and nonconsent rates per comparison and compared the way this information was extracted. Some studies did not report all the information we needed (frequencies of consent and nonconsent). We contacted the primary authors of these studies, but when we were unable to retrieve the correct data, we excluded the study from the meta-analysis. This was the case for two studies (one for wording, one for sponsorship). In addition, one study on sponsorship was deleted because the network model would not run because of inconsistent variances.

We calculated the risk difference per treatment comparison with the metafor package version 2.4-0 (Viechtbauer, 2010). Afterwards, we performed the network meta analyses with the netmeta package version 1.2-1 (Rücker, Krahn, König, Efthimiou & Schwarzer, 2020). All analyses were performed in R version 4.0.2 (R Core Team, 2020).

2.4 Results

2.4.1 Experiments Addressed in Previous Studies

The inventory of all conducted experiments in the 45 eligible manuscripts shows that there is a large variation in the aspects of the consent question covered. We summarized the experiments per experimental category; see Table 2.1 for the full inventory.

Out of the 45 studies, there were two types of experiments that were conducted in at least 10 studies: question wording and sponsor. We therefore decided to perform a network meta-analysis for these aspects only.

Table 2.1 Inventory of all experiments conducted in the studies.

Study	Sponsor	Wording		Consent Form Appearance	Control	Data Release	Data type	Incentive	Interviewer Appearance	Mode	Opt-in/-out	Position	Purpose of Data Use	Study Duration	Text Length	Topic	Type of Participants
Al Baghal 2016																	■
Al Baghal 2019										■							■
Antommaria 2018						■											
Balestra 2016				■													
Becker, 2020		■		■													
Berry 2013											■						
Beuthner *in press*		■					■	■				■					
Bhatia 2018													■				
Boyd 2015				■													
Breuer 2019								■									
Brelsford, 2019															■		
Briscoe, 2020	■																
Bryant 2006																	
Burstein 2014						■											■
Critchley 2015	■					■											
Das 2014										■					■		
Edwards *in press*															■		
Eisnecker 2017												■					
Fobia 2019													■				
Grande 2015	■												■				■
Graves 2019								■									■
Halevi 2015	■							■									
Jäckle *in press*										■							
Keusch 2019	■				■			■						■		■	
Kim 2015	■												■				
Kim 2017	■						■										
Kreuter 2016																	
Kreuter 2018		■															
McGuire 2011						■											
Nodora 2017	■																
Pascale 2011		■									■						
Passmore 2020									■				■				
Peycheva *in press*										■							
Pratap, 2019	■																
Sakshaug 2013		■										■					
Sakshaug 2014		■															
Sakshaug 2015		■										■					
Sakshaug 2019a		■															
Sakshaug 2019b		■								■		■					■
Sala 2014												■					
Sanderson 2017						■											
Shah 2018	■																
Struminskaya 2020a	■	■			■												
Struminskaya 2020b		■			■												
Weydert 2019								■									

2.4.2 Meta-analytical Models

2.4.2.1 Sponsorship

We performed a network meta-analysis on nine studies, comparing the effect of different sponsors on data linkage consent rates. The four possible sponsors are academia (e.g., universities or university hospitals), government (e.g., National Statistical Institutes (NSIs) or federal agencies), companies (e.g., a market research company or pharmaceutical company) and nonprofit organizations (e.g., hospitals or patient organizations). Figure 2.2 presents the general outcomes of the meta-analysis in a forest plot.

The I^2 statistic describes the percentage of observed heterogeneity that would not be expected by chance. Heterogeneity (I^2) is 100% and thus very large. Therefore, we look at the random effects meta-analytical model.

Study sponsors affect consent rates. Sponsorship by a university evokes similar consent rates for data linkage as sponsorship by a nonprofit organization. Sponsorship by a university yields 1.5 times higher consent rates than sponsorship by a governmental organization. Sponsorship by a company evokes the lowest consent rates, three times lower than sponsorship by a university or nonprofit organization.

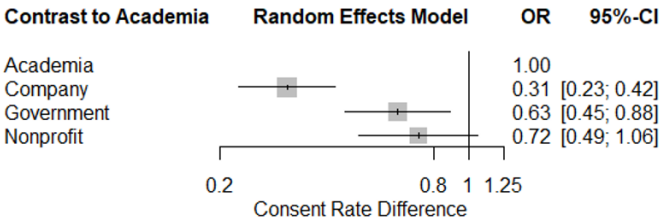


Figure 2.2 Forest plot for sponsorship, random effects model, with academia as the reference category.

2.4.2.2 Wording

We also performed a network meta-analysis on ten studies comparing the effect of different wording on data linkage consent rates. The five wording conditions are neutral framing (no loss or benefit is mentioned), emphasizing the benefit of data linkage (benefit framing) in terms of cost *("To reduce costs")*, result *("To improve on the accuracy of our results")*, time *("To keep the interview as brief as possible")* and loss *("Survey data will be "less useful" if consent is not provided")*.

The I^2 statistic describes the percentage of observed heterogeneity that would not be expected by chance. Heterogeneity (I^2) is 99.2% and thus very large. Therefore, we look at

the random effects meta-analytical model. Figure 2.3 presents the general outcomes of the meta-analysis in a forest plot.

Wording affects consent rates; emphasizing benefits or losses increases consent rates for all conditions. We find a significant increase for benefit framing in terms of costs or time savings and loss framing. Loss framing and benefit framing in terms of costs yield 1.3 times higher consent rates than neutral framing. Emphasizing the decrease in time it takes to fill out the survey or complete the interview increases consent rates 1.1 times compared to not emphasizing a reason to consent. The effect of framing in terms of costs is tested in only one study, however, leading to a very large confidence interval for *cost*.

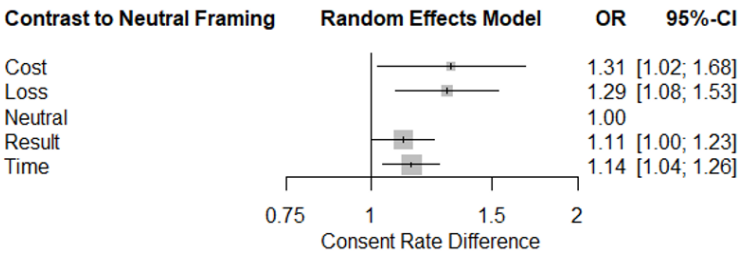


Figure 2.3 Forest plot for wording, random effects model, with neutral framing as reference category.

2.4.3 Descriptive Review

In the following section, all experimental categories that were covered by fewer than ten studies will be reviewed. See Table 2.1 for the full inventory of experiments per aspect of the consent question covered. These aspects are discussed in alphabetical order.

2.4.3.1 Consent Form Appearance

Three studies experimented with consent form appearance (Balestra et al., 2016; Becker et al., 2020; Boyd et al., 2015).

Balestra et al. (2016) presented participants with an online consent form for a personal genomics study. Individuals were randomly assigned to view the consent from positive-, negative- or mixed-valence comments beside the text of the consent form. The positive-valence condition, for example, included a comment box with *"I like that they've thought about data protection at every step of the process"* in the margin of the screen. Consent rates did not differ across the valence conditions. Participants who spent less time studying the consent form were more likely to consent when they were exposed to positive-valence comments. These findings are likely explained by participants not taking much time to

think about the consent question.

Becker et al. (2020) experimented with different argument strength levels: logical arguments, illogical arguments or no arguments. Argument strength has a significant effect on consent rates; no arguments yield the lowest response rates, and logical arguments yield the highest response rates. Surprisingly, giving illogical arguments is more effective than giving no arguments.

Boyd et al. (2015) compared a professionally designed version (of the information pack and consent form) with a standard version for administrative record linkage. Both versions contained the same elements, text and photographs, but the standard version looked plain in comparison to the designed pack. They found no differences in consent rates between the two versions.

2.4.3.2 Control

Three studies investigate the effect of giving respondents the ability to control data collection (Keusch et al, 2019; Struminskaya et al, 2020a; Struminskaya et al., 2020b).

The findings are mixed. Struminskaya et al. (2020a; 2020b) specified that respondents would be able to view what information they send to the research institute and are able to undo the measurement later if they wish to do so. Struminskaya et al. (2020b) found that giving respondents control increased respondents' hypothetical willingness to share sensor (location) data but not for actually sharing the data. Struminskaya et al. (2020a) found no effect of giving control on hypothetical willingness. Keusch et al. (2019) found that giving respondents the option to switch off sensor data collection increased respondents' hypothetical willingness to participate in passive data collection.

2.4.3.3 Data Release

Five studies experimented with data release options (Antommaria et al., 2018; Burstein et al., 2014; Critchley et al., 2015; McGuire et al., 2011; Sanderson et al., 2017). All studies asked respondents to participate in a biobank, where biological samples from respondents are linked to environmental and medical information. These databases may be accessible to anyone online (public release) or only to qualified researchers (restricted release). Antommaria et al. (2018), Critchley et al. (2014) and Sanderson et al. (2017) found no differences between public or private release for the general population and for participants who sought care at a network site.

In the study of McGuire et al. (2011), participants could choose between public release, restricted release [accessible only to approved researchers] or no release. Consent rates for restricted release are comparable to those for private release. Approximately 85% consent

to data sharing, but one-third of the entire sample chose restricted release. Burstein et al. (2014) compared parents of pediatric patients and adult patients on the same three data sharing preferences as McGuire et al. (2011). They also found that most participants ultimately consented to broad public release (73.5% and 90.3%). However, parents were more restrictive in their data release decisions regarding their children, as they are more concerned about unknown future risks.

2.4.3.4 Data Type

For the data type for which linkage is sought, we identified administrative data, social media data, biodata and passive data. We included only two studies that varied this type of data experimentally. There are more studies that ask for different data sources (e.g., health and administrative records), but when these sources fall within the same category, they are not taken into account here.

Two studies asked for consent to link different types of data (Beuthner et al, in press; Kim et al., 2017). Kim et al. (2017) asked for biodata (blood and tissue samples) and administrative data (health test results). They found no large difference in willingness to share these data types. Beuthner et al. (in press) asked for all four data types and found similar consent rates (approximately 30% on average) for all data types, except for consent rates for sharing administrative banking data, which was approximately 18%.

2.4.3.5 Incentive

Six studies experimented with incentives (Beuthner et al., in press; Breuer et al., 2019; Graves et al., (2019); Halevi et al., 2015; Keusch et al., 2019; Weydert et al., 2019). These studies experimented with the amount, timing, or type of incentive.

The findings on the effect of the amount of the incentive on consent to rates are mixed. Halevi et al. (2015) repeatedly asked respondents about their willingness for different incentive amounts and plotted a trendline between consent rate and incentive amount. The trendline for people who are willing to share their information is very steep between 5 and 50 dollars and flattens above 50 dollars. Keusch et al. (2019) compared no incentive, 10 euros (either for downloading an app or at the end of the study) and two times 10 euros (for downloading and for completing the study). No incentive yielded much lower consent rates (19.5%) than one incentive of 10 euros (37.1% and 38.3%). Giving 10 euros twice yields the highest consent rates (46.1%). In contrast, Beuthner et al. (in press) found no effect between an incentive of 1 or 3 Euro per consent decision. However, 1 and 3 euros are both rather low incentive amounts, which may influence these results. Weydert et al. (2019) investigated the effect of an incentive on data sharing with an internet broker. They compared no incentive with a 60 dollar incentive and found lower consent rates when offering an incentive. Their idea is that generous monetary compensation shows

respondents the monetary value of their personal data and potential privacy protection loss, of which they may otherwise have been unaware. In other words, the incentive increases the salience of privacy concerns.

Keusch et al. (2019) and Breuer et al. (2019) both found no effect for the timing of the incentive. Keusch et al. (2019) compared the effect of two conditional incentives: an incentive of 10 euros for downloading an app at the beginning of the study with an incentive of 10 euros at the end of the study. Breuer et al. (2019) compared the effect of an unconditional (prepaid) and conditional (postpaid) incentive of 5 euros.

Graves et al. (2019) compared two types of incentives: a lottery ("chance to win a gift voucher") or an item ("very fashionable designer leggings"). Women who were offered the item were more likely to consent than those who had a chance to win a gift voucher.

2.4.3.6 Interviewer Appearance

Passmore et al. (2019) investigated the effect of ethnicity or interviewer appearance on consent rates among African Americans. They found higher willingness for Black or "other" ethnicity compared to a white interviewer.

2.4.3.7 Mode

Five studies compared different interview modes (Al Baghal, 2019; Das et al., 2014; Jäckle et al., *in press*; Pecheva et al., *in press;* Sakshaug et al., 2019b) for the consent question.

Overall, interviewer-based modes yield higher consent rates than self-administered modes (Al Baghal, 2019; Jäckle et al., *in press;* Peycheva et al., *in press;* Sakshaug et al., 2019b). Peycheva[3] et al. *(in press)* also differentiated between telephone and face-to-face surveys and concluded that face-to-face surveys yielded even higher consent rates than telephone surveys. Das et al. (2014) experimented with asking for consent by email or letter and found slightly higher opt-out rates for the email condition.

2.4.3.8 Opt-in/Opt-out

Two studies experimented with opt-in and opt-out conditions (Berry et al., 2013; Pascale et al., 2011). Opt-out yielded much higher consent rates in both studies (21.3% vs 95.6% and 83.9% vs 98.3%). According to Berry et al. (2013), attrition and participation rates did not differ between opt-in or opt-out consent.

[3] Peycheva *(in press)* used a self-selection design, so selection biases cannot be excluded.

2.4.3.9 Position

Six studies investigated the position of the consent question (Beuthner et al., *in press;* Eisnecker et al., 2017; Sakshaug et al., 2013; Sakshaug et al., 2015; Sakshaug et al., 2019b; Sala et al., 2014). These studies all looked at different aspects of the position.

Beuthner et al. *(in press)* investigated the consent rates by position (one to seven), with seven consecutive consent rates. Respondents did not know beforehand how many data linkage requests they would see. Consent rates drop after the first consent question, but consent rates are nearly consistent between the second and seventh positions. They found this effect throughout the four different data types we discern: administrative, sensor data, social media data and biodata.

Sakshaug et al. (2013; 2015; 2019b) and Sala et al. (2014) investigated the effect of position in the questionnaire. Sala et al. (2014) found a positive effect for positioning the consent question in context compared to placing the question at the end of the questionnaire. Respondents were asked for consent to link administrative data (e.g., national insurance contributions, benefits and tax records, savings and pensions) directly after a module of questions that asked about the receipt of state benefits and other payments. Sakshaug et al. (2013) and Sakshaug et al. (2019b) found a positive effect of positioning the consent question at the beginning of the interview or survey compared to the end of the interview or survey. Sakshaug et al. (2015) found no significant effect between beginning-and-end positioning. They did, however, find an interaction effect with loss/benefit framing; respondents in the gain-framing condition consent to linkage at a higher rate than those in the loss-framing condition when response usefulness was emphasized for responses to subsequent survey items.

Eisnecker et al. (2016) examined the position of the linkage request in different waves of a panel study. They found no difference in consent rates between wave 1 and wave 2. They found lower consent rates for respondents who dropped out in the wave after the consent request.

2.4.3.10 Purpose of Data Use

Five studies experimented with the purpose of the data requested in the consent question; e.g., what the data would be used for (Bhatia & Breaux, 2018; Fobia et al., 2019; Grande et al., 2015; Kim et al., 2015; Passmore et al., 2020). Bhatia and Breaux (2018) found that participants were most willing to share their information for purposes that they perceive to be more beneficial to society. Passmore et al. also found that a research goal in line with participant values resulted in higher percentages of willingness to share biodata. In their case, *"to help people with a disease I care about"* yielded higher consent rates than usage in a biobank and a government-sponsored research center.

Grande et al. (2015) differentiated between linking electronic health information for research, quality improvement and marketing. Willingness to share data for marketing uses was generally lower than for the other two uses. Respondents were most willing to share data for research purposes. Kim (2015) investigated the likelihood of respondents to consent to share data for research and for healthcare. Almost half of respondents were likely to consent to both healthcare and research uses of their electronic data, but fewer would consent to healthcare than to research. Fobia et al. (2019) examined four different question frames related to evidence-based policy making; respondents were less likely to consent to data linkage when the purpose was described as being for informed decision-making than for efficient use of taxpayer money, government accountability, or community beliefs.

2.4.3.11 Study Duration

Keusch et al. (2019) experimented with the survey duration or the time (or duration) that the app would run and collect passive data. Participants were less willing to participate in passive mobile data collection with a longer duration of data collection (one month versus six months).

2.4.3.12 Text Length

Three studies experimented with the length of the consent text, a short version and a long version (Brelsford et al., 2019; Das et al., 2014; Edwards & Briddle, *in press*).

Das et al. (2014) and Brelsford et al. (2019) found slightly higher consent rates for the longer text version, whereas Edwards & Briddle *(in press)* found slightly higher consent rates for the shorter version. In the study of Brelsford et al. (2019), more respondents felt the short or simplified form contained the right amount of information compared to the traditional form. The differences between the long and short versions are, however, not significant (or not tested (Brelsford et al., 2019)) in these studies.

2.4.3.13 Topic

Keusch et al. (2019) experimented with the topic of the survey for which data are collected (mobility, consumer behavior or social interaction). Topic had no effect on respondents' willingness to share passive data.

2.4.3.14 Type of Participants

Five studies compared different participant groups. Two studies compared consent rates between different (naturally occurring) participant groups (Burstein et al., 2014; Grande et al., 2015), one study compared consent of mothers for themselves and for their children (Al Baghal et al., 2016), one study compared consent rates within the general population with a panel (Al Baghal et al., 2019), and one study compared consent rates of several (open)

recruitment methods (Graves et al., 2019).

Burstein et al. (2014) compared data sharing preferences between parents of pediatric patients and adult patients. Most parents (73.5%) and participants (90.3%) consented to broad public release. Parents were more concerned about future risks for their child and therefore more restrictive in their data sharing decisions. Grande et al. (2015) compared participants with and without a prior diagnosis of cancer on their willingness to share their electronic health information. Overall, their willingness was similar, and cancer patients were even more willing to share (sensitive) genetic information.

Al Baghal et al. (2016) compared consent rates to administrative data when mothers were asked to consent for themselves and for their children. The large majority of mothers (88.2%) gave the same answer to the consent request for themselves and for their children. Most mothers that did not give the same answer to both requests, only gave consent for themselves.

Al Baghal et al. (2019) compared consent rates to link Twitter data from two panels (NatCen Panel, Innovation panel) with a cross-sectional survey within the general population (British Social Attitudes [BSA] 2015). No differences in consent rates were found between panel and nonpanel members.

Graves et al. (2019) conducted open recruitment using a variety of methods: Facebook (advertisements), other web activities (such as Twitter, Instagram, YouTube), referrals, traditional media and fashion promotion. Consent rates to administrative record linkage differed by method of recruitment, with women who were recruited via Facebook being less likely to provide consent, while women who were recruited via fashion promotion were more likely to consent.

2.5 Discussion

In this study, we performed a systematic literature review to present an overview of the literature on consent to data linkage. We included 45 studies that investigated different aspects of the consent question. We aggregated and summarized the literature on the consent question and discovered several aspects of the consent question that affect consent rates. Earlier studies found that consent rates varied. This research aims to help improve the number of people who consent to data linkage in future research.

Our results on wording show that providing respondents with reasons for giving consent increases consent probabilities. Respondents seemed most sensitive to question wording, which would imply that their survey data lost value if the data were not linked, as loss

framing yielded high consent rates with a relatively small confidence interval. Mentioning that data linkage would result in a decrease in respondent burden or a reduction of survey costs also positively affects consent rates. It appears to be most effective to give an argument for data sharing compared to not giving an argument; the exact framing has less impact on the consent rates. Becker et al. (2020) also showed that even giving illogical arguments for data sharing yields higher consent rates than giving no arguments. For future research, giving logical arguments for participation is thus advised over not mentioning any reason for participation.

Our results on sponsors show that studies by academia or nonprofit organizations yield the highest consent rates compared to governmental organizations or commercial companies. In particular, companies yield very low consent rates. The study sponsor probably evokes higher or lower trust. For future research, it seems effective to make the organization more (for universities or nonprofit organizations) or less (for companies) salient in the consent request. Another solution for companies might be to collaborate with universities and nonprofit organizations and emphasize their logo in the consent request.

When looking at the descriptive review, some aspects of the consent question to data linkage affect consent rates. However, some of these aspects are only investigated in one or a few studies, which makes it difficult to draw strong conclusions. The first aspect that affects consent rates is the mode of the consent question; interviews yield higher consent rates than self-administered surveys, especially if the interviewer is more like the respondent. The second aspect is the incentive; higher incentives yield higher consent rates. This effect levels off after a certain amount of money (50 dollar). Above that amount, incentives may even decrease respondents' willingness. High incentive amounts may make the value of the survey data more salient to respondents, which may in turn increase worries about privacy. Surprisingly, the studies were inconclusive about the difference between conditional and unconditional incentives, whereas in the survey literature, unconditional incentives are seen as the best option to prevent nonresponse (Göritz, 2006; Pforr, 2016; Singer et al., 1999). The mechanisms behind the decision to give consent are thus probably different than the mechanisms behind the decision to respond to a survey. The third aspect is the position of the consent question within the survey. Asking for consent at the beginning of a survey or "in context" is more effective than asking for consent at the end of a survey. A study on panel recruitment also found higher recruitment rates for beginning positioning, but these higher rates followed higher attrition rates compared to end positioning (Toepoel, 2016). Therefore, the position of the panel request ultimately did not matter for the size of the panel. In addition, when respondents are asked for consent repeatedly, consent rates drop after the first question regardless of the data being asked for. The fourth aspect is the study duration; respondents are less likely to give consent when the passive data collection period is longer. The last aspect that affects

consent rates is opt-out. Opt-out is more effective than opt-in; however, opt-out is no longer allowed since the General Data Protection Regulation (GDPR) is employed.

Another important finding is the effect of the relevance of the research project to the respondent on consent rates. We conjecture that when a research project is more relevant, trustworthy or important to the respondents, respondents are more likely to give consent to data linkage. This is, for example, visible for purpose, where respondents are more likely to give consent when the purpose feels more beneficial for society or for themselves. We also see that cancer patients are more likely to give consent to cancer-related research than the general population.

We see no clear effect of data type (administrative data, biodata, social media data or passive data), data release options (public or private release), participants (naturally occurring groups, panels or recruitment methods), text length (longer or shorter versions of the consent form), giving respondents control (to see or change their data), or study topic. The results of experiments conducted around these aspects are insignificant or vary in direction.

Unfortunately, for many aspects, the number of included studies was insufficient to perform a meta-analysis. The availability of more studies would increase the usefulness and insights for these aspects. The field of data linkage is relatively new, and the number of studies is increasing rapidly, so this problem may be solved in the near future. The number of included studies in the meta-analyses was sufficient but not very high. We accounted for the high heterogeneity by interpreting the random effects model instead of the fixed effects model (Schwarzer, Carpenter, & Rücker, 2015).

Another limitation was the variation in the scales used to assess consent rates. Some studies asked for hypothetical willingness on a binary scale and some on a continuous scale, while some asked for actual consent. All these studies and scales were transformed into odds ratios and used in our meta-analysis, which may have increased the heterogeneity of the meta-analytical model.

In the future, an increasing amount of data will have great value in linking results from attitudinal surveys to behavioral or factual information from other sources (registers, sensors, social media, etc.). This study showed that there is a relatively small body of evidence for how we should ask study participants for consent to data linkage. The results from this meta-analysis show that how researchers ask consent questions matters greatly for how high consent rates are achieved. However, more research on ways to frame the consent request is needed, especially on how to combine elements from the request that in this study have been studied as separate aspects. For example, if consent rates

increase even further when an interviewer states reasons for participation, or is one aspect sufficient, or does this mechanism only work for self-administered surveys. The results from this study can be useful in designing future consent protocols.

Chapter 3

Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process?

This chapter is published as: Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process?. *Survey Research Methods,* 13, 195-213.

Author contributions: The Netherlands Institute for Social Research and CentERdata designed the study and collected the data. AE, PL and VT contributed to the concept and analysis plan of the paper. AE performed the statistical analyses. AE wrote the paper, PL and VT critically reviewed the paper.

Abstract

The increasing use of smartphones opens up opportunities for novel ways of survey data collection, but also poses new challenges. Collecting more and different types of data means that studies can become increasingly intrusive. We risk over-asking participants, leading to nonresponse. This study documents nonresponse and nonresponse bias in a smartphone-only version of the Dutch Time Use Survey (TUS). Respondents from the Dutch LISS panel were asked to perform five sets of tasks to complete the whole TUS: 1) accept an invitation to participate in the study and install an app, 2) fill out a questionnaire on the web, 3) participate in the smartphone time use diary on their smartphone, 4) answer pop-up questions and 5) give permission to record sensor data (GPS locations and call data). Results show that 42.9% of invited panel members responded positively to the invitation to participate in a smartphone survey. However, only 28.9% of these willing panel members completed all stages of the study. Predictors of nonresponse are somewhat different at every stage. In addition, respondents who complete all smartphone tasks are different from groups who do not participate at some or any stage of the study. By using data collected in previous waves we show that nonresponse leads to nonresponse bias in estimates of time use. We conclude by discussing implications for using smartphone apps in survey research.

3.1 Introduction

The increasing popularity of smartphones opens up opportunities for novel ways of data collection in survey research (e.g. Miller, 2012) that could complement and partly substitute survey questions. Unlike Internet (browser) surveys, smartphone apps enable the collection of auxiliary data, such as GPS locations or communication behavior through mobile phones (Dufau et al., 2011; Miller, 2012; Raento, Oulasvirta, Eagle, 2009). Smartphones incorporate a large number of sensors (e.g. accelerometers, GPS, light and proximity sensors) which can be logged passively, providing a large and detailed set of measurements about respondents and their environment (Cottrill et al., 2013; Ermes, Parkka, Mantyjarvi, Korhonen, 2008).

Most traditional surveys face declining response rates (De Leeuw, Hox, & Luiten, 2018). Respondents are increasingly reluctant to participate (Groves & Heeringa, 2006) especially when surveys are long and questions burdensome. Galesic (2006) shows that the more burdensome questions are, the less motivated respondents are to answer them. Smartphone surveys can get increasingly burdensome and intrusive as we ask respondents to share more personal information. Respondents may not be willing to share these kinds of data due to privacy concerns (Revilla, Couper, & Ochoa, 2019). Although smartphones offer great possibilities for better measurement, we risk overasking our participants.

One type of survey research that faces this trade-off between measurement quality and response rates is time use research. Response rates in time use surveys, traditionally conducted with paper diary studies, are generally not very high (Abraham, Maitland, & Bianchi, 2006; Knulst & Van den Broek, 1999; Stoop, 2007). For example, the response rate for the Dutch Time Use Survey (TUS) ranges between 18% in 1995 and 40.3% in 2011–2012 (Cloïn et al., 2013; Statistics Netherlands (CBS), 2013; Van Ingen, Stoop, & Breedveld, 2008). Response rates for time use studies in the United States are 54.6% (Abraham et al., 2006) and 45% in the United Kingdom (Fisher & Gershuny, 2013). These findings hold in other types of diary studies: An American web-based dietary study had a response rate of 10% and a diary completion rate of only 7.4% (Thompson et al., 2014).

Diary studies are burdensome. Moreover, time use data based on paper diaries suffer from measurement error and recall problems. Measurement in diary studies could potentially improve when conducted through an app (Sonck & Fernee, 2013).

The primary research objective of our study was to investigate the effect of asking intrusive questions through a smartphone app study on survey response rates, predictors of nonresponse and nonresponse bias in a smartphone TUS conducted in the Dutch probability-based LISS Panel. Panel members were asked to participate in several tasks

varying in intrusiveness and burden (completing surveys, a diary, sharing sensor data and answering pop-up questions). Nonresponse bias could be introduced or accumulate at every step. We will use attributes of the participants (e.g. personality, demographics, smartphone familiarity) to predict nonresponse in these different stages. Subsequently we will examine whether this nonresponse influences our survey estimates, resulting in nonresponse bias.

3.2 Theoretical background

3.2.1 Decision Making Process for Survey Response

In order to understand why respondents do or do not participate in the separate parts of the smartphone study, we use the leverage-saliency theory (Groves, Singer, & Corning, 2000). According to this theory, respondents make a decision to participate or not with every request to participate. In making this decision, different respondents place different importance on factors of the survey request. One respondent might value the topic of a survey, another the incentive offered, or the emphasis that the advance letter puts on value for society. Negative leverage factors could be survey burden, privacy concerns, or topic difficulty. Someone's propensity to participate depends on the number of positive and negative factors perceived in the request (leverage) and the relative importance to the respondent (saliency) (Groves et al., 2000; Keusch, 2015).

Most research testing the leverage-saliency theory use experiments to vary aspects of the leverage and saliency in the survey request explicitly. In our study, leverage and saliency were varied more naturally as respondents were asked to perform tasks that are different in nature. For example, worries about privacy may influence the willingness to share GPS data, whereas pop-up questions that interrupt daily life may annoy some participants (e.g. those who are busy). The diary study in our app is the most time-consuming part of the study, so respondents who are sensitive to burden may dropout in this task. Because the nature of the tasks in our study differ, participants may be willing to participate in one task of the TUS, but not in another. This difference in willingness to participate per task may then cause nonresponse bias to vary per task as well.

The leverage-saliency model was developed with a one-time decision in mind, such as in cross-sectional surveys. In many smartphone studies, participants have to make multiple decisions to participate in different tasks which are not independent. From longitudinal surveys, we know that once a panel member has agreed to participate in a study, this decision is likely to be followed by continuous participation (Lemay, 2010). Similarly, once a respondent has not participated in a task, he or she may be more likely to not participate in subsequent tasks either. We can incorporate this longitudinal feature by including prior

decision as a separate factor in our leverage-saliency model.

3.2.2 Time Use Surveys

The self-completed time use diary is considered to be the most reliable and accurate data collection instrument to obtain information on the activity patterns of participants (Michelson, 2005). Many European time use surveys follow the Harmonized European Time Use Survey (HETUS) guidelines (Eurostat, 2009). Time diaries are also used in other fields, for example to measure physical activity (e.g. Bouchard's Physical Activity Record, BAR; see Bouchard et al., 1983) or dietary intake (e.g. Automated Self-Administered 24-hour, ASA24 Dietary Assessment Tool; see Subar et al., 2012), but how the diary is designed varies between studies. Most time diary studies cover the full 24 hours of a day and divide the day into ten-minute timeslots. This works as a cognitive cue and reduces omissions due to forgetfulness (Belli, Shay, & Stafford, 2001). Time diaries also allow the collection of contextual information such as whom the respondent was with, and typically make a distinction between main and side activities. Time use diaries sometimes only span a short time (e.g. one day only), but ideally cover a longer period so that infrequent activities are also captured (Gershuny, 2012).

Time use diaries present challenges for data collection. First, diaries are burdensome to complete, often resulting in response rates that are lower than those of one-time questionnaire-based surveys. Second, if respondents do not complete the diary regularly throughout the day, recall problems may arise resulting in less accurate data. Third, the administration costs are high because the manual coding and entering of data from paper is very labor intensive (Minnen et al., 2014).

Conducting time use research on a smartphone could create a more user-friendly and less burdensome instrument compared to the traditional paper-based TUS. Respondents can complete the diary any time of the day, as long as they have their smartphone with them. In contrast to the paper diary which is usually left at home and filled out at the end of the day, using a smartphone app diary makes it possible to remind respondents to fill out their diary several times per day. This may help to reduce the recall-problem (Lai et al., 2010). In addition, smartphones enable the collection of auxiliary data, such as GPS locations or communication behaviors (Raento et al., 2009), which can reduce the number of questions we need to ask. Finally, a smartphone app can significantly reduce the time and costs of data processing. An app enables the use of pre-coded categories of time use, avoiding coding efforts after data collection has been completed.

There are also some potential pitfalls of using smartphones for diary studies. Coverage error may lead to bias when certain groups or members of the population who do not

own a smartphone are automatically excluded. Participants may further be unwilling or insufficiently able to use an app, leading to nonresponse and nonresponse bias. A pilot study of a smartphone TUS by Chatzitheochari et al. (2018) showed promising results regarding response and data quality. 97 cohort members of the UK Millennium Cohort Study were invited for the pilot and could self-select into the web (28%), or a smartphone version (41%) of the TUS. The paper diary (20%) was only offered to participants without a personal computer or smartphone, or who refused to use the web and smartphone modes. There was a nonresponse rate of 11%. Mode choice was similar by gender and household income. Results show that the completion rate for the smartphone (48% on day 1 and 30% on day 2) and web version (33% and 30%) were slightly lower than the paper version (63%). Comparisons of the measurement quality across modes found that there were fewer item-missings in the smartphone app mode and more contextual data (location, and who the respondent was with). Due to a very small sample size and a non-randomized mixed mode design, the results from Chatzitheochari et al. (2018) can however only be treated as indicative.

3.2.3 Analytical Framework and Hypotheses

Apart from factors specific to the prediction of nonresponse in our smartphone TUS, there are also general predictors of nonresponse relevant to our study. Demographic characteristics such as gender, age, educational level, occupation, ethnicity, household status and size, urbanicity, and marital status have been shown to generally correlate with response propensity in surveys (Fan & Yan, 2010; Groves, Cialdini, & Couper, 1992). Certain demographic variables decrease the contact likelihood, such as urbanicity, living alone, and living without children (Abraham et al., 2006; Groves, 2006). Other demographic characteristics tend to increase refusal rates, such as ethnicity, educational level and age (Lipps, 2009; Lugtig, 2014; Van Ingen et al., 2008). Often there is no clear theoretical link between sociodemographic variables and nonresponse (Fan & Yan, 2010). They can however be important to include as predictors in order to compare studies, and to investigate nonresponse bias.

In longitudinal studies, socio-psychological variables are thought to be more closely related to why participants keep participating or drop out from a study. For example, respondents who are more “agreeable” on the Big Five personality scale are more cooperative, whereas “conscientious” people tend to be more reliable and determined (Costa & McCrae, 1992). Both should lead to a higher commitment to the survey and have been associated with lower dropout rates (Lugtig, 2014; Richter, Körtner, & Saßenroth, 2014). In contrast, people with high levels of “extraversion” are reported to become easily bored and distracted (Costa & McCrae, 1992), leading to dropout. “Openness” also seems to have a robust effect on response propensity as people high in openness are considered

to be more interested in new experiences and intellectually curious (Richter et al., 2014; Salthouse, 2014).

Respondents' survey attitude is also an important indicator for survey commitment (De Leeuw, Hox, Silber, Struminskaya, & Vis, 2019; Stocké, 2006). Respondents with a positive survey attitude—who think surveys are important and enjoy answering them—are less likely to attrite (Stocké, 2006). These respondents may place less importance on the burden of the survey, and more on the value of the survey.

Other predictors that we will use in this study are specific to our smartphone TUS: we expect that respondents who use their smartphone frequently are more willing to use this device for survey completion (De Bruijne & Wijnant, 2014; Mavletova, 2013) and that privacy concerns will prevent respondents from sharing GPS data.

In our analytical models to study nonresponse we will include sociodemographic, socio-psychological variables, survey attitudes and specific predictors for each task in order to study whether the correlates of nonresponse differ across the different tasks of the TUS. Following the leverage-saliency model, we expect the correlates to differ per task as the aspects of the request are also different. For example, worries about privacy may particularly influence the willingness to share GPS data, whereas busyness may be mainly related to interruptive pop-up questions, and survey burden to the most time-consuming part, the time use diary. Apart from looking at correlates of nonresponse, we will also look at bias in estimates of time use. As the smartphone TUS was conducted within an existing panel, we had some basic information on time use available for almost all sample members. We will test whether biases on these variables cancel each other out over the different stages, or reinforce each other.

3.3 Methods

3.3.1 Sample

In this study we used data from the LISS (Longitudinal Internet Studies for the Social sciences) panel, administered by CentERdata (Tilburg University, The Netherlands). The LISS panel started in 2007 and is the principal component of the project Measurement and Experimentation in the Social Sciences (MESS). The LISS panel consists of about 8000 individuals who complete online questionnaires every month. These questionnaires cover a large variety of domains including work, income, housing, time use, political views, values and personality. For more information about the LISS panel, see, Scherpenzeel and Das (2010).

The panel is based on a simple random sample of households drawn from the Dutch population register by Statistics Netherlands and aims to be representative of the Dutch population (Scherpenzeel, 2011; Scherpenzeel & Das, 2010). After the first sample was drawn in 2007, a refreshment sample was recruited between June and December 2009. Non-Internet households that could otherwise not participate are provided with a computer and Internet connection. Using the response metrics of Callegaro and Disogra (2008) the initial recruitment rate for the LISS panel was 63% and the profile rate 48% (Scherpenzeel, 2009). Retention is about 90% a year (Toepoel, 2013). See Table 3.4 in the Appendix for the demographical composition of the LISS panel. In 2012 and 2013, the smartphone TUS was administered to the LISS panel.

3.3.2 The Smartphone Time Use Survey

To study nonresponse and nonresponse bias, we examine response patterns in the Dutch Time Use Survey (TUS), developed and coordinated by the Netherlands Institute for Social Research (NL: Sociaal en Cultureel Planbureau [SCP]). In 2012 the SCP first conducted a small-scale test of their TUS through an app on a smartphone (Sonck & Fernee, 2013), after which the app was adapted and administered to the LISS panel.

Respondents in the TUS without a smartphone, or who preferred using a phone provided by LISS, could borrow one from the LISS panel: 52% of the participants in our study (n = 1120) used a borrowed phone. Respondents received an incentive of 15 euros for every hour of participating. The research process of the TUS in the LISS panel consisted of several stages, listed here in chronological order:

1. **Willing to participate.** Participants of the Dutch LISS panel were asked in three different surveys of the LISS panel—conducted in August 2012, March 2013 and July 2013—whether they were interested in participating in future smartphone surveys. Those who said yes to any of these three requests were considered for the TUS. See Table 3.4 in the Appendix for the demographics of the TUS sample.

2. **Time Use Survey.** To minimize possible seasonal influences on the time use data, data were collected for an entire year. Data collection started in September 2012. After this, each month a different batch of 176 panel members was invited (Sonck & Fernee, 2013). This resulted in a sample of 2154 participants in September 2013. People in this sample were invited for every stage of the TUS, even if they did not participate in the prior stage(s).

3. **Pre-questionnaire.** Participants started by completing a pre-questionnaire on the web that mimicked the normally used paper diary TUS. This pre-questionnaire contained more than 200 questions on various topics, like smartphone use, feeling in control about one's life, reasons to work certain hours, social support, childcare, hobbies and household composition.

4. **Diary.** About a week after the invitation to the pre-questionnaire, participants were asked to download an app, and complete a diary in which they recorded their activities on two randomly selected days; one weekday and one weekend day. The activities were predefined from a list of 41 categories following HETUS guidelines (Eurostat, 2009). Participants had to complete the full 24 hours (from 04:00 am to 04:00 am the next day) using ten-minute intervals. See Figure 3.1 for a screenshot of the app and time use diary. The left panel shows the screen where activities were reported by a hypothetical respondent, the middle panel shows an overview of the respondent's set of recorded activities, the right panel shows the questions that were asked using Experience Sampling (see 4. Pop-up questions below). The app was available for iOS and Android users.

   When participants failed to complete one of the two days they were assigned to, they were invited to participate on a third day. This third day was exactly one week after the first weekday. We coded participants as respondents in this task when they filled out the diary for at least one day.

5. **Pop-up questions.** On the same days as the diary, participants received six pop-up questions which were sent at random times of the day between 8:00 am and 10:00 pm. The pop-up question would show up on the screen for ten minutes. After this ten-minute interval the question disappeared and could not be answered anymore to ensure real-time feelings were measured. These pop-up questions asked respondents either about their emotional state or smartphone use in the past hour. See Figure 3.1 for a screenshot of the pop-up questions. We coded participants as respondents in this task when they answered at least one pop-up question.

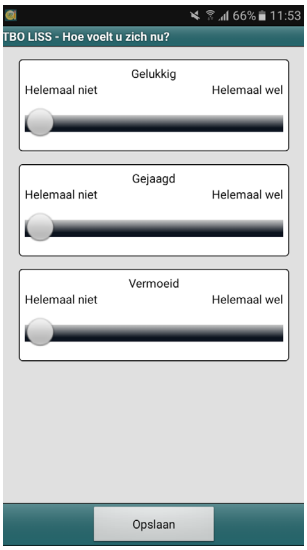
6. **Sensor Data.** Participants were asked permission to passively record additional data through the app. These data included communication

data (number of incoming and outgoing calls and text messages) and GPS locations. When downloading the app, participants were asked permission for this passive data collection. By default, the GPS tracker was turned on, but respondents could turn this off any time. We could not directly observe who turned off the GPS tracker and when this happened; we can only observe the number of GPS data points available per participant. The distribution of GPS data points showed a clear peak around 576–600 GPS data points for a two-day period implying that under normal conditions one location measurement was taken every five minutes. Participants with fewer than 576 data points were assumed to have turned off their GPS tracker or phone at some point during the study and were treated as nonrespondents in this stage. We performed sensitivity analyses treating everyone with at least 278 GPS point as respondents, but found no differences with the results we present below.

7. **Post-Questionnaire.** After all smartphone tasks were completed, respondents completed another questionnaire on the web. This post-questionnaire had the same design as the pre-questionnaire and contained about 180 questions on various topics, like Internet use, use of social networks, and family life.

(a) The screen where activities were reported. TUS LISS—Your activity is: add main activity. You did this between: 12:00 and 12:10. Were you alone of with someone you know? Alone/ With children up to 9 years old / With other family members / With someone else you know. The buttons below in the screen show “Copy previous activity”; and “Save”.

(b) The day overview of one set of recorded activities. TUS LISS—Day overview Wednesday 26 July.
Time & Activities:
05:30 Sleeping, 05:40 Sleeping, 05:50 Sleeping, 06:00 Sleeping, 06:10 Sleeping, 06:20 Sleeping, 06:30 Sleeping, 06:40 Sleeping, 06:50 Sleeping, 07:00 Eat-ing/ Drinking at home, at work/ school,
07:10 Eating/Drinking at home, at work/school, 07:20 Personal or Med-ical care, 7:30–7:40 Personal or Med-ical care. The button below in the screen shows “Add activity”.



(c) The three pop-up questions, ask-ing: TUS LISS—How do you feel at this moment? Happy, Rushed, Tired. The scale labels are: Not at all—Extremely. The button below in the screen shows “save”.

Figure 3.1 Screenshots op the TUS app.

3.3.3 Instruments

We used a variety of background variables from previous waves of the LISS panel to predict nonresponse. As most of these variables are measured annually, we used the data from respondents that were recorded closest before the start of the TUS.

Sociodemographic characteristics. We used a set of sociodemographic characteristics: gender, age, net income, highest level of education (7 categories), and number of children living with the respondent.

Personality. Five personality factors were computed: openness, conscientiousness, extraversion, agreeableness and neuroticism. These five factors are based on the Big Five, a taxonomy for describing the basic dimensions of personality (Costa & McCrae, 1992). We used the self-rating Big Five Questionnaire (Goldberg et al., 2006). This questionnaire consists of fifty items on which respondents must rate how they apply to them on a five-point scale. See the Appendix, Table 3.5 – 3.8, for all question wordings and results of our factor analyses.

Survey attitude. The LISS panel contained nine questions about one's general attitude towards surveys. These items asked the participants for example whether they think surveys are important for society, and exhaustive to answer. Three factors for survey attitude were computed; survey enjoyment, survey value and survey burden (De Leeuw et al., 2019).

Privacy. Two factors regarding privacy concerns were computed; trust and worries. The factor trust covers three questions about how much participants trust different organizations to keep their personal information private. The factor worries covers two questions about how worried participants are about their privacy. These questions were part of a survey conducted in July and August 2008. For that reason, data were not available for all respondents.

Smartphone use. The factor smartphone use is based on questions in the pre-questionnaire of the TUS. Participants were asked whether they used their smartphone for 22 internet activities, for example for watching television, surfing the web and sending tweets. We calculated a factor score based on these 22 activities, see the Appendix Table 3.5 – 3.8. Participants were also asked to report for themselves how often they used mobile Internet on their phone. The correlation between this measure and the factor 'smartphone use' was $r = 0.733$.

Participation history. We calculated the proportion of surveys panel members participated in relative to the number of invitations they received over the course of participation in the LISS Panel.

Prior decision. Prior decision is a stage-specific measure of continuous participation. We added participation (0 = no, 1 = yes) in the previous stage of the smartphone TUS as a predictor for the subsequent stages. For example, participation in the pre-questionnaire is the prior decision for participation in the TUS diary.

3.3.4 Missing Data

Missing data on covariates were imputed using the EM-algorithm in the Missing Values Analysis module in SPSS 24.0 (IBM Corp., 2016). 14.87% of the cases were complete. There were 77 different missing data patterns. By far the largest group had missing data only on variables that made up the factor score "privacy". 44.5% of the cases had a missing value on privacy. Sociodemographic variables were available for everyone, except for income (5.7%), urbanization (1.1%) and educational level (1.0%).

3.3.5 Analyses

Our first, descriptive objective was to see how many participants participate in every stage, how many drop out and how many return. Second, we predicted nonresponse in every stage of the smartphone TUS, using the covariates described in the section above. We ran multivariate, logistic regression models with response (0 = nonresponse, 1 = response) as the dependent variable. We also included participation history and prior decision in our logistic regression analyses hierarchically to investigate the effect of continuous participation. Third, we investigated nonresponse bias to discover if and how nonresponse influences the survey estimates.

All models were estimated using R 3.4.1 (R Core Team, 2020). We conducted several multivariate logistic regression analyses and calculated Average Marginal Effects (AME) with the R-Package "mfx" (Fernihough, 2014). Marginal effects (MFX) are the estimated probabilities that the respondent participates for a specific, marginal change in the explanatory variable, holding all other variables fixed. AME expresses the average MFX of the explanatory variable on the dependent variable (Mood, 2010). We report AME instead of odds-ratios because odds-ratios reflect unobserved heterogeneity (Mood, 2010). Unobserved heterogeneity is the variation in the dependent variable that is caused by variables that are not observed and thus not included as predictors in the model. As this unobserved heterogeneity varies across models, we cannot simply compare the effect of specific predictors at different stages. AME's are not affected by unobserved heterogeneity

and can thus be compared across models and groups. We transformed the AME's into percentages in the tables we present, to make them easier to interpret.

3.4 Results

Figure 3.2 shows the responses at different stages of the TUS. Only participants who said that they were willing to participate were invited for stage 1, the pre-questionnaire. The reported numbers in the squares represent the participants who start in that specific stage.

At three moments in 2012 and 2013, 7296 participants of the LISS panel received the question whether they would be willing to participate in a smartphone survey. Participants who answered they were not willing to participate were treated as refusals, while participants who did not answer this question at all were treated as noncontacts. Following the AAPOR 2006 guidelines, we observe a noncontact rate of 16.0% ($n = 1168$), and a refusal rate of 41.1% ($n = 2996$). 42.9% of all respondents ($n = 3132$) said they would be willing to participate in our smartphone study (Callegaro & Disogra, 2008).

In Figure 3.2, the numbers in the arrows represent the response probabilities for the subsequent stages, conditional on whether the respondent participated in the preceding stage. The arrows between stages represent consistent participation. For example, 85.6% of the participants who complete the pre-questionnaire also complete the diary, and 74.0% of the participants who complete the diary share GPS data. The arrows that do not connect boxes represent participants returning after missing a previous stage. For example, 32.0% of the nonrespondents in the pre-questionnaire fill out the diary, and 0.2% of the nonrespondents in the diary share GPS data.

The flowchart (Figure 3.2) shows that at every stage participants drop out. This results in a smaller number of participants at every subsequent stage. An exception forms the stage of GPS sharing, which has the lowest number of participants ($n = 1193$). A relatively large group misses one stage, but then returns to complete the next stage. This is especially so for the diary and post-questionnaire.

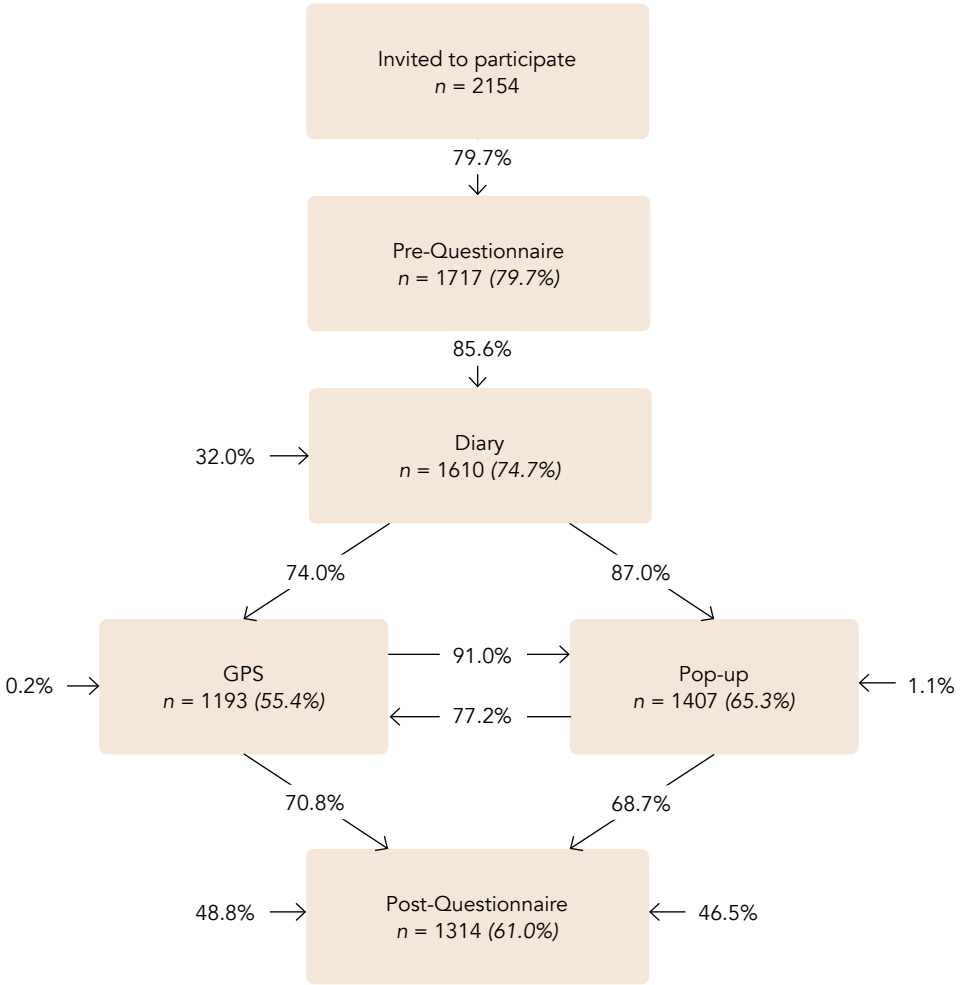


Figure 3.2 Flowchart of response behaviour in the smartphone TUS. The percentages indicate the response probability per stage, dependent on participation (arrow from one stage to another) or no participation (arrow from the outside to the square) in the preceding stage.

3.4.1 Willing to Participate

In order to identify who is willing to participate in the smartphone survey, we ran a multivariate logistic regression model. In this analysis we excluded the noncontacts[4] at the invitation stage since we are uncertain whether they would be willing to participate.

[4] When comparing the sociodemographic characteristics of the contacts and noncontacts it appeared that noncontacts are on average younger and more likely to live in a household with children.

Table 3.1 shows the result of this multivariate logistic regression model (e.g. 0 = not willing, 1 = willing). Age, educational level, extraversion, conscientiousness, openness, survey value, enjoyment and burden, worries about privacy, and smartphone ownership are all significant predictors of being willing to participate in smartphone studies. For example, if people own a smartphone, their probability of participating increases by 21.5% conditional on the other covariates in the model. Furthermore, a one-year increase in participant age decreases the probability to participate by 0.55%. Participation history, added to the model to investigate the additional effect of respondents' prior commitment to LISS panel, does not have a significant effect. The results of the model without controlling for participation history are reported in Table 3.9 in the Appendix.

Table 3.1 Average Marginal Effects for Predicting Willingness to Participate.

	AME	**Std. Err.**
Sociodemographics		
Gender	-1.84	1.57
Age	-0.55***	0.06
Educational level	3.74***	0.51
Number of children	1.04	0.69
Income	-0.02	0.02
Personality		
Neuroticism	0.07	0.76
Extraversion	-2.08**	0.76
Agreeableness	-0.91	0.84
Conscientiousness	-4.41***	0.77
Openness	3.42***	0.77
Survey Attitude		
Survey value	3.81***	1.15
Survey enjoyment	10.96***	1.21
Survey burden	-2.62*	1.17
Privacy		
Trust	1.41	0.95
Worries	-3.69***	0.85
Smartphone Use		
Smartphone Ownership	21.51***	1.88
Continuous Participation		
Participation History	4.33	3.97
Nagelkerke R2	.190	

Note. p < .001 '=***' . < .01 ='**'. < .05 = '*'.

3.4.2 Participation in the Time Use Survey

To assess who does and who does not participate in the different stages of the smartphone TUS, we ran separate multivariate logistic regression analyses per stage (0 = not willing, 1 = willing). To investigate the effect of continuous participation and to control for effects of the previous stage, we added prior decision and participation history next to sociodemographic, socio-psychological and smartphone specific predictors in our model. The results of these final multivariate logistic regression analyses are presented in Table 3.2. The results of the models without controlling for continuous participation are reported in Table 3.10 in the Appendix.

First, we look at the sociodemographic predictors across all stages. Sociodemographic variables have no effect on participating in the pre-questionnaire. Willingness to participate in the smartphone parts of the study declines with age. The willingness to fill out the diary increases 1.54% per educational level. All other effects of sociodemographic variables across the stages are nonsignificant and/or small.

For the socio-psychological variables we find no large effects. Willingness to participate in the pre-questionnaire is higher for respondents who are more conscientious. With every increase of one standard deviation in the factor score conscientiousness the probability to participate increases by 1.95%. Willingness to share GPS data is larger for respondents who are more introvert.

Furthermore, smartphone use significantly predicts sharing GPS data. Participants who use their phone more often are more likely to be willing to share their GPS data. No variables related to survey attitude or privacy have an effect in any of the stages.
Finally, we look at the effects of continuous participation. Respondents prior commitment to the LISS panel, measured by participation history, increases respondents' willingness to fill out the pre-questionnaire or diary. Participating in the diary seems a strong predictor for the two subsequent phases. Participants for the diary have an 86.69% higher probability of being willing to answer pop-up questions, and 72.97% to share GPS data. Prior decision does not have a large effect on the post-questionnaire, but filling out the diary increases the willingness with 19.54% and sharing GPS data with 8.86%.

The explanatory power of our final model is particularly high for the smartphone parts of the study. The high explained variance for the pop-up (0.706) and GPS stage (0.554), in combination with the large effect of participating in the diary, implies that (non)response in these stages is mostly conditional on the previous stage.

Table 3.2 Average Marginal Effects for Participants' Willingness to Participate in the Different Stages.

	Pre-Questionnaire		Diary		Pop-up		GPS		Post-Questionnaire	
	AME	Std. Err.	AME	Std. Err.	AME	Std. Err.	AME	Std. Err.	AME	Std. Err.
Sociodemographics										
Gender	-1.14	2.02	-3.10	2.27	-4.24	4.42	4.44	3.00	1.25	2.58
Age	-0.06	0.07	-0.41***	0.08	-0.81***	0.16	-0.06	0.11	0.08	0.09
Educational level	-0.46	0.69	1.54*	0.78	-0.32	1.46	-0.52	1.00	1.05	0.88
Number of kids	0.28	0.76	0.83	0.90	0.16	1.74	-1.65	1.13	-0.63	0.99
Income	0.19	0.11	0.12	0.12	0.42	0.25	0.21	0.17	0.10	0.13
Personality										
Neuroticism	0.36	0.95	0.44	1.07	-0.92	1.99	-2.02	1.38	0.59	1.21
Extraversion	-1.67	0.92	-1.15	1.05	0.37	1.96	-4.05**	1.38	-0.79	1.17
Agreeableness	1.05	1.01	1.63	1.12	-0.61	2.16	-1.12	1.47	1.44	1.28
Conscientiousness	1.95*	0.91	1.43	1.05	-1.02	1.99	2.37	1.34	0.82	1.19
Openness	-1.13	0.93	-1.47	1.05	-0.45	1.97	0.89	1.33	-0.78	1.19
Survey Attitude										
Survey value	-0.09	1.46	2.21	1.60	3.94	3.13	1.11	2.12	0.15	1.84
Survey enjoyment	-0.36	1.50	-1.91	1.67	-7.40*	3.17	-3.94	2.18	0.90	1.89
Survey burden	-0.54	1.52	-2.01	1.69	-0.19	3.08	-1.00	2.16	-0.67	1.95
Privacy										
Trust	0.53	1.17	1.40	1.30	-1.09	2.37	1.47	1.65	0.92	1.48
Worries	-0.69	1.03	1.94	1.11	0.42	2.07	-0.02	1.45	0.48	1.28
Smartphone Use										
Smartphone use	-0.18	1.09	0.97	1.34	3.26	2.33	4.25**	1.57	-2.65	1.41
Continuous Participation										
Participation History	33.14***	4.29	13.04*	5.34	10.87	11.83	6.90	7.21	8.32	6.25
PD: Pre-Questionnaire	-	-	53.84***	2.54	10.45	6.37	10.89**	3.71	1.45	3.14
PD: Diary	-	-	-	-	86.69***	1.05	72.97***	1.23	19.54***	4.56
PD: Pop-up	-	-	-	-	-	-	-	-	3.15	3.86
PD: GPS	-	-	-	-	-	-	-	-	8.86**	2.93
Nagelkerke R2	.077		.326		.706		.554		.120	

Note. PD = "Prior decision". p < .001 = '=***' . < .01 = '**'. < .05 = '*'.

3.4.3 Nonresponse Bias

Apart from looking at the characteristics of those who respond and not respond, the final goal of this paper was to see how nonresponse matters for substantive statistics. The goal of the TUS is to estimate time use. To investigate the bias in these estimates we compared the time use of respondents and nonrespondents. We derived our estimates of time use for both groups from the "Social Integration and Leisure" and "Work and Schooling" questionnaires of the LISS panel. The "Social Integration and Leisure" study was conducted in February/March, and "Work and Schooling" in April/May 2012, roughly a year before the TUS. Respondents' time use could have changed in the meantime, but we find it safe to assume that over the whole sample on average no shifts occurred. We excluded respondents for whom data were missing (n = 618). For our analyses of nonresponse bias, we looked at four distinctive groups of respondents:

1. Respondents who said they were not willing to participate ($n = 2601$).
2. Respondents who said they were willing to participate, but never did ($n = 130$).
3. Respondents who participated in the TUS, but only in the pre- and/or post- web questionnaire, not in any smartphone parts (pop-up, diary, GPS) ($n = 278$).
4. Respondents who participated in all tasks ($n = 622$).

Table 3.3 shows how many hours per week respondents on average spend on several activities. These activities are working, watching television, doing volunteer work, doing sports, going out (to bars, restaurants, cinema), making music, going to the theater (ballet, plays or musical) and doing creative things. In addition, we calculated the absolute and relative nonresponse bias.

An analysis of variance (ANOVA) test revealed that there are no differences between the groups in time spent on volunteer work, music, doing sports, theater or creative activities. The groups do differ significantly on time spent working ($F(3, 3627) = 18.49$, $p < .001$, $\eta_p^2 = .015$), watching television ($F(3, 3627) = 15.25$, $p < .001$, $\eta_p^2 = .012$), and going out ($F(3, 3627) = 4.60$, $p = .003$, $\eta_p^2 = .004$).[5]

The relative and absolute bias of watching television and working are also high, 3.70 (18.3%) and 4.96 hours (30.1%). The relative bias of theater and creativity are also high (47.8%), but this is due to very low prevalence of these activity in general. The absolute bias is only 0.04 and 0.08 hours, which is less than five minutes.

[5] When we included all cases, also those including missings on some items of time use from earlier LISS surveys used to assess nonresponse bias, most results did not differ. The only difference is that group 1 and 3 now differ significantly on doing sports ($F(3, 3930) = 2.95$, $p = .031$, $\eta_p^2 = .002$). By excluding incomplete cases, we excluded $n = 148$ for group 1, $n = 17$ for group 2, $n = 29$ for group 3 and $n = 35$ for group 4.

Post-hoc analyses reveal that only certain groups differ. First, full respondents who participated in all tasks of the TUS, appear to work significantly more hours than participants of the other three groups. The groups with (partial) nonrespondents do not differ from each other in the time spent on working. Second, full respondents watch significantly less television than the group that was not willing to participate at all (group 1) and the group that participated only in the non-smartphone parts (group 3). Finally, the full respondents (group 4) and respondents who were not willing to participate (group 1) go out less often than the other two groups (group 2 and 3).

Table 3.3 Average Time Spent on Several Activities by Four Distinctive Groups of (Non)respondents, and the Absolute and Relative Bias Induced by Nonresponse.

	Not willing	Willing, no participation	Participation in non-smartphone parts	Full Respondents	Sample Mean	Absolute Bias (hours)	Relative Bias (in %)
Work	15.35	15.38	16.27	21.42	16.46	4.96	30.13
Volunteer Work	1.84	1.29	1.66	1.68	1.78	0.10	5.63
Watching TV	21.22	19.21	19.94	16.56	20.26	3.70	18.26
Sports	1.95	2.38	2.47	2.07	2.03	0.04	1.97
Going out	1.20	1.61	1.73	1.12	1.24	0.12	9.68
Music	0.12	0.03	0.18	0.11	0.11	0.00	0.00
Theater	0.11	0.04	0.09	0.05	0.09	0.04	47.78
Creativity	0.58	0.60	0.88	0.51	0.59	0.08	13.53

Note. The absolute and relative bias are calculated by comparing the Full Respondents (respondents) to the Sample Mean (full sample), as described by Groves and Peycheva (2008).

3.5 Discussion and Conclusion

This paper shows how nonresponse differs over different tasks in a smartphone survey. This study is a first step into building a methodological framework for understanding nonresponse error in smartphone surveys. We know little about whether people are willing to perform smartphone-specific survey tasks and how this affects data quality.

Results show that 42.9% of the LISS panelists are willing to participate, and that from this group of willing respondents 74.6% actually complete the smartphone TUS. Only 28.9% of the willing respondents complete all the tasks of the study. The basic response rate is comparable to other, offline time use surveys, although we do not take nonresponse in the

recruitment of LISS into account (Abraham et al., 2006; Van Ingen et al., 2008).

Predictors of nonresponse differ per task. However, some variables consistently predict nonresponse in every stage. Being younger, more conscientious, more open and more introvert increase the response probability of participating in every task in the smartphone TUS. These results are consistent with other, offline Time Use Studies (e.g. Abraham et al., 2006; Stoop et al., 2005; Van Ingen et al., 2008) and other longitudinal studies (Costa & McCrae, 1992; Lugtig, 2014; Richter et al., 2014). Smartphone ownership is an important predictor for being willing to participate in a smartphone study. Even though respondents could borrow a smartphone to participate, many participants were unwilling to do so. Haan, Lugtig and Toepoel (2019) showed that device familiarity is an important predictor for using a particular device to complete a survey. This is confirmed in the actual TUS, where participation in some of the smartphone parts is predicted by age, rather than smartphone use. We conjecture that the respondent's attitude towards and familiarity with the specific functions of the smartphones are the main determinants of willingness to participate, rather than actual frequency of smartphone use, or age. Providing equipment is probably not enough to warrant participation in smartphone studies. In this study, the LISS panel tried to increase familiarity by showing an instruction manual and video. This may be a possible way to improve participation, but it is not very powerful as many participants did not view this video. Using interviewers might be a more powerful tool to increase device familiarity or to ensure everyone sees the video.

When panel members participated in the prior stage of the TUS, they are more likely to participate again. This finding can be explained by the foot-in-the-door technique (Cialdini, 1993), where a small initial request increases compliance with the next, larger request. Our study was rather successful in this respect, probably because it started with a regular survey respondents were used to.

When we frame our results in light of the leverage-saliency model we see that respondents who have a more positive smartphone attitude are more likely to participate in the smartphone parts of the study than in the survey parts. This suggests that respondents make a thoughtful decision to participate. Most of the variance is however explained by continuous participation, not by aspects of the survey request of that specific task. Future research could test the leverage-saliency model more extensively, by varying the survey request wording per task.

Nonresponse in itself may not be a problem, as long as it does not influence the survey estimates (Groves, 2006). However, in this study, nonresponse does influence the survey estimates and therefore induces nonresponse bias. A specific group participates in the smartphone parts; this group works more and watches less TV than the average LISS panel

participants. The respondents who only complete the pre- or post-questionnaires are more similar to those who do not participate in our study at all than like the full respondents. These results also replicate the results of Van Ingen et al. (2008) and Abraham et al. (2006), who also found that busy people are more likely to participate in an offline TUS. According to Stoop (2005) busy or working people are more involved in society. This involvement may lead to a higher probability to participate, but probably also leads to a more positive smartphone attitude that might be work-related. Future research could shed more light on this relation and the occurrence of nonresponse bias specifically in the smartphone parts.

Unfortunately, we were not able to see directly who turned off GPS tracking. We coded participants with few GPS data points as nonrespondents as we assumed they turned off their GPS tracker. However, it is also possible that these people had their phones switched off during data collection, and only turned them on to complete the diary. Other causes may be technical problems, or empty batteries. Since it is difficult to pinpoint what the reasons are for the missing GPS data, it is difficult to predict how this introduces bias in the estimates. Future research should make this clearer.

A further limitation of our study is that we used respondents who were already participating in the LISS panel. The advantage of using a panel is the large amount of auxiliary variables available for predicting nonresponse and studying nonresponse bias. However, the experienced sample may be generally more willing to participate in surveys, and smartphone surveys in particular. The response rate in our study was comparable to earlier, offline time use surveys. If we take into account though the fact that LISS respondents are used to doing research, it is likely that repeating our smartphone study in an independent cross-sectional sample would result in a lower response rate than found in TUS conducted with paper diaries (Abraham et al., 2006; Van Ingen et al., 2008).

There is a long way to go before we can use smartphones as the sole data collection mode in general population studies. This study proves that we can conduct smartphone-app surveys with some success. Future research should focus on how to increase smartphone familiarity and on ways to convince people to do survey related tasks on smartphones. Consent surveys could shed more light into these issues. Smartphone surveys are promising tools for social research, but if we want to improve response rates and decrease response bias, there is still a lot of work to do on easing respondents into a task that is at least to some degree intrusive.

3.6 Appendix

Table 3.4 Demographical Composition of the LISS panel, Time Use Survey sample and the Dutch Population.

		LISS panel[c]	TUS sample	Dutch Population[b]
Age				
	Average (in years)	40.6[a] (21.98)	44.38a (17.09)	41.6[a]
	65+	15.9	14.0	16.2
Gender				
	Male	49.0	46.4	49.5
	Female	51.0	53.6	50.5
Ethnicity				
	Native Dutch Background	87.2	87.6	79.1
	Migration Background	12.8	12.4	20.9
	Urbanicity			
	Extremely Urban	12.7	13.5	20.5
	Very Urban	25.3	25.4	24.0
	Moderately Urban	23.9	23.2	18.1
	Slightly Urban	22.3	23.0	18.6
	Not Urban	15.9	15.0	18.8
Composition (per household)				
	Single HH	28.3	17.4	36.8
	Couple with children	34.5	44.3	27.3
	Couple without children	30.2	30.4	29.2
	Single with children	5.6	6.7	6.8
	Other	1.4	1.2	0.6

[a] Average age reported as mean instead of percentages. [b] Dutch population statistics correspond to individually based statistics, with January 1st 2012 as the base date. Statistics can be found on http://statline.cbs.nl (Statistics Netherlands, 2012). [c] LISS panel composition of September 2012.

Table 3.5 Factors Sociopsychological Variables.

	Factor				
	1	**2**	**3**	**4**	**5**
Neuroticism					
Get stressed out easily.	.70	-.02	.05	-.09	-.12
Am relaxed most of the time.	-.61	-.01	.01	.19	.16
Worry about things.	.67	.06	.09	-.21	-.01
Seldom feel blue.	-.52	-.06	.07	.08	.06
Am easily disturbed.	.73	-.04	.08	-.13	-.09
Have a soft heart.	.39	.00	.26	.00	.01
Get upset easily.	.75	-.08	.07	.05	-.08
Change my mood a lot.	.69	.03	-.02	.14	.07
Have frequent mood swings.	.68	.03	-.02	.13	.04
Get irritated easily.	.54	-.03	-.11	-.05	.07
Often feel blue.	.73	.01	-.05	.00	.04
Extraversion					
Am the life of the party.	.12	.64	.01	.01	.15
Don't talk a lot.	-.07	-.67	-.07	.05	.10
Feel comfortable around people.	-.06	.48	.23	.00	.01
Keep in the background.	.01	-.79	.22	.06	.08
Start conversations.	.00	.67	.16	-.06	-.02
Have little to say.	.17	-.51	-.13	.11	-.01
Talk to a lot of different people at parties.	.03	.63	.22	.12	-.05
Don't like to draw attention to myself.	.00	-.62	.34	-.05	-.07
Don't mind being the center of attention.	.01	.57	-.11	.14	.17
Am quiet around strangers.	.14	-.64	-.08	.03	.11
Agreeableness					
Feel little concern for others.	.02	-.03	-.51	.12	.06
Insult people.	.18	.09	-.30	.20	.22
Sympathize with others' feelings.	.09	-.18	.88	.12	-.01
Am not interested in other people's problems.	.02	-.02	-.65	-.03	.01
Am not really interested in others.	.09	-.06	-.68	-.01	.05
Feel others' emotions.	.13	-.04	.69	.15	.14
Make people feel at ease.	.11	.31	.38	-.06	.06
Take time out for others.	.05	.03	.66	-.02	.04
Am interested in people.	-.04	.06	.71	.11	.04

Conscientiousness					
Am always prepared.	.03	.03	-.09	-.51	.09
Leave my belongings around.	.00	.02	.14	.71	.11
Make a mess of things.	.30	-.06	.01	.61	.10
Get chores done right away.	.07	.10	-.01	-.55	-.05
Like order.	.26	.00	-.10	-.76	.05
Often forget to put things back in their proper place.	.08	.05	.04	.61	.02
Shirk my duties.	.26	.06	-.09	.48	-.07
Follow a schedule.	.25	-.03	.03	-.52	.13
Spend time reflecting on things.	.09	-.12	.10	-.31	.29
Am exacting in my work.	.16	.00	-.03	-.36	.29
Openness					
Have difficulty understanding abstract ideas.	.28	.02	.02	.05	-.43
Pay attention to details.	.12	-.03	.13	-.29	.37
Have a vivid imagination.	.11	.15	-.08	.17	.40
Am not interested in abstract ideas.	.17	.00	-.05	.02	-.36
Have a rich vocabulary.	-.10	-.02	.09	-.01	.48
Have excellent ideas.	-.08	.04	-.02	-.11	.56
Do not have a good imagination.	.25	.00	-.07	.13	-.28
Am quick to understand things.	-.17	-.13	.07	-.11	.57
Use difficult words.	.00	-.07	-.09	.18	.55
Am full of ideas.	.05	.19	.02	-.08	.54

Note. Standardized factor loadings. We used an EFA with Promax Rotation in IBM SPSS 24. The Cumulative Explained Variance is 38.53%. Factor correlations range between -.285 and .370.

Table 3.6 Factors Survey Attitude.

	Factor		
	1	**2**	**3**
Survey Value			
Surveys are important for society.	.80	-.01	.19
A lot can be learned from information collected through surveys.	.79	-.01	.13
Survey Burden			
Completing surveys is a waste of time.	-.40	.51	.06
I receive far too many requests to participate in surveys.	.06	.55	-.05
Opinion polls are an invasion of privacy.	-.13	.55	.11
It is exhaustive to answer so many questions in a survey.	.15	.53	-.24
Survey Enjoyment			
I really enjoy responding to questionnaires through the mail or Internet.	.12	-.05	.80
I really enjoy being interviewed for a survey.	.07	-.02	.56
Surveys are interesting in themselves.	.48	-.04	.50

Note. Standardized factor loadings. We used an EFA with Promax Rotation in IBM SPSS 24. The Cumulative Explained Variance is 54.26%. Factor correlations range between -.327 and .424.

Table 3.7 Factors Privacy.

	Factor	
	1	**2**
Trust		
How much do you trust each of the following to keep the information they collect from you confidential: public opinion research companies	1.01	.05
How much do you trust each of the following to keep the information they collect from you confidential: market research companies	.63	-.03
How much do you trust each of the following to keep the information they collect from you confidential: government agencies, like Statistics Netherlands	.49	-.06
Worries		
In general, how worried are you about your personal privacy?	.01	.76
Different private and public organizations have personal information about us. How concerned are you about whether or not they keep this information confidential?	-.05	.81

Note. Standardized factor loadings. We used an EFA with Promax Rotation in IBM SPSS 24. The Cumulative Explained Variance is 58.65%. Factor correlation is -.230.

Table 3.8 Factor Smartphone Usage.

	Factor
Please indicate whether you ever use a mobile phone for...	1
Watching television	.42
Watching films online	.73
Listening to the radio	.46
Listening to your own music	.59
Reading news sites and daily newspaper	.66
Reading magazines	.46
Playing online games	.46
Playing offline games	.48
Emailing	.71
Reading other people's twitter messages	.47
Sending your own twitter messages	.41
Visiting social media network sites	.68
Visiting online forums or discussion groups	.34
Sending short text messages via the Internet	.69
Telephoning via the Internet or making video calls	.39
Downloading music or video files	.49
Uploading videos, photos or music	.56
Online banking	.57
Shopping or ordering goods via the Internet	.48
For navigation services	.65
To search specific information on the Internet	.75
Just to surf around on the Internet	.64

Note. Standardized factor loadings. We used an EFA with Promax Rotation in IBM SPSS 24. The Cumulative Explained Variance is 31.56%.

Table 3.9 Average Marginal Effects for Predicting Willingness to Participate without Participation History.

	AME	Std. Err.
Sociodemographics		
Gender	-1.82***	1.57
Age	-0.54***	0.05
Educational level	3.83	0.50
Number of children	0.98	0.68
Income	-0.02	0.02
Personality		
Neuroticism	0.04	0.76
Extraversion	-2.14**	0.76
Agreeableness	-0.97	0.84
Conscientiousness	-4.29***	0.76
Openness	3.34***	0.76
Survey Attitude		
Survey value	3.81***	1.14
Survey enjoyment	11.05***	1.21
Survey burden	-2.65*	1.17
Privacy		
Trust	1.47	0.95
Worries	-3.69***	0.85
Smartphone Use		
Smartphone Ownership	21.33***	1.87
Nagelkerke R2	.190	

Note. p < .001 = '=***' . < .01 ='**'. < .05 = '*'.

Table 3.10 Average Marginal Effects for Participants' Willingness to Participate in the Different Stages Without Prior Decision.

	Pre-Questionnaire		Diary		Pop-up		GPS		Post-Questionnaire	
	AME	Std. Err.	AME	Std. Err.	AME	Std. Err.	AME	Std. Err.	AME	Std. Err.
Sociodemographics										
Gender	-1.38	2.02	-3.40	2.21	-4.21	2.47	0.55	2.58	0.31	2.51
Age	0.05	0.07	-0.28***	0.08	-0.49***	0.09	-0.21	0.09	0.01	0.09
Educational level	0.59	0.68	1.94**	0.74	1.71*	0.82	1.20	0.87	1.71*	0.84
Number of kids	-0.05	0.77	0.61	0.86	0.56	0.96	-0.70	0.99	-0.58	0.96
Income	0.13	0.11	0.16	0.12	0.31*	0.14	0.26	0.14	0.14	0.13
Personality										
Neuroticism	-0.02	0.94	0.26	1.03	-0.06	1.14	-1.06	1.20	0.46	1.16
Extraversion	-2.09*	0.91	-2.17*	1.01	-1.90	1.11	-4.57***	1.17	-1.69	1.13
Agreeableness	0.69	1.00	1.70	1.09	1.17	1.21	0.53	1.28	1.69	1.24
Conscientiousness	3.04***	0.90	3.20**	0.99	2.79*	1.11	4.27***	1.17	2.11	1.13
Openness	-2.03*	0.92	-2.60**	1.01	-2.59*	1.12	-1.43	1.17	-1.64	1.14
Survey Attitude										
Survey value	0.04	1.43	1.87	1.53	2.83	1.73	2.00	1.84	0.79	1.77
Survey enjoyment	0.78	1.46	-0.87	1.58	-3.13	1.77	-3.30	1.87	0.55	1.81
Survey burden	-0.65	1.51	-1.97	1.61	-1.84	1.81	-2.44	1.92	-1.29	1.87
Privacy										
Trust	0.91	1.15	1.80	1.24	1.22	1.38	2.34	1.46	1.55	1.42
Worries	-0.78	1.02	1.25	1.08	1.28	1.20	1.03	1.28	0.80	1.24
Smartphone Use										
Smartphone use	-0.51	1.08	0.26	1.20	1.37	1.34	3.33*	1.40	-2.18	1.35
Nagelkerke R2	0.035		0.041		0.055		0.044		0.027	

Note. p < .001 = '=***' . < .01 ='**'. < .05 = '*'.

Chapter 4

Where You at? Using GPS Locations in an Electronic Time Use Diary Study to Derive Functional Locations.

This chapter is published as: Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, S., & Toepoel, V. (2019). Where You at? Using GPS Locations in an Electronic Time Use Diary Study to Derive Functional Locations. *Social Science Computer Review.*

Author contributions: WB and SR designed the study and organized the data collection and database. CentERdata programmed the app and hosted the data collection. AE, PL, and VT designed the data preparation and analyses method. AE prepared the data and performed the statistical analyses. AE wrote the paper. WB, PL, SR, and VT critically reviewed the paper.

Abstract

Smartphones enable passive collection of sensor data alongside survey participation. Location data add context to people's reports about their time use. In addition, linking global positioning system data to self-reported time use surveys (TUSs) can be valuable for understanding how people spend their time. This article investigates whether and how passive collection of *geographical* locations (coordinates) proves useful for deriving respondents' *functional* locations. Participants of the ongoing Children of Immigrants Longitudinal Survey in the Netherlands were invited to participate in a TUS administered with a smartphone app that also unobtrusively tracked respondents' locations. Respondents reported their activities per ten-minute interval in a smartphone diary app (n = 1,339) and shared their geographical location data (n =1,264). The correspondence between the functional locations derived from the time use data and those derived from the geographical location data was assessed by calculating the percentage of intervals in which both measures are similar. Overall, results show that home locations can be automatically assigned reliably but that respondent information is required to reliably assign work or school locations. In addition, location tracking data contain many measurement errors, making it difficult to record valid locations. Multilevel models show that the variability in correct classifications is intrapersonal and largely predicted by phone type, which determines location measurement frequency.

4.1 Introduction

Smartphones are increasingly viewed as groundbreaking data collection tools for studying human behavior (Link et al., 2014; Miller, 2012; Raento, Oulasvirta, & Eagle, 2009). In the Netherlands, 90.3% of the total population owned a smartphone with Internet access in 2018, with percentages over 98.5 for individuals between 12 and 45 years (Statline, 2019). People carry smartphones around naturally and use them continuously for information and communication purposes in everyday life. This allows researchers to directly interact with respondents at any time and at any location (Raento et al., 2009). Because most people are used to sharing opinions and information through the apps on their mobile phones, they probably see smartphone-based data collection tools as natural extensions of the many functions of their phones, which could lower the threshold to use it for research.

Smartphones have sensors that researchers could use to collect high-intensity data passively, that is, without the measurements requiring any respondent activity beyond giving a one-time permission to share sensor measures with the researchers. We do not know whether respondents are willing to do give such permissions and neither do we know how useful sensor data actually are for social scientific research purposes.

A growing body of research on sensor data (e.g., Anhoj & Moldrup, 2004; Chatzitheochari et al., 2017; Plowman & Stevenson, 2012; Sonck & Fernee, 2013) illustrates the new types of data that can be collected via mobile phones. Notwithstanding their importance, many of them are small case studies and pilot projects (e.g., Cottrill et al., 2013; Sugie, 2018). Questions remain regarding the usefulness and data quality of sensor data.

This article addresses methodological challenges of analyzing and integrating survey and sensor data. More specifically, our purpose is to assess whether in time use surveys (TUSs) administered on smartphones, the passive recording of participants' geographical locations (the coordinates of the locations) is sufficient to establish their functional locations (the natural functions of the locations). In other words, if we know a person's current activity and geographic coordinates, do we also know the function of their current location, for example, whether they are at home, at school, or somewhere else? As location tracking data incorporate information about people's time use, linking global positioning system (GPS) location data to self-reported TUSs can potentially be a beneficial practice. The main aims of this article are to propose a method for analyzing GPS data, to integrate location and survey data, and to explain variability. If we understand how to use location tracking data, and if locations are measured accurately, we may be able to automatically record functional locations in TUSs in the future.

4.2 Background

Traditional and paper diary studies are commonly used by researchers to measure a respondent's time use, travel behavior, physical activity, or dietary intake. Respondents are asked to take detailed notes about all their activities for several days. The accuracy of the data depends on respondents' memory and the effort they are willing to put in filling out the diary. Due to the high burden, respondents may be less willing to put effort in the diary over time and may also be inclined to postpone filling out their diaries leading to higher recall bias. Analyses of two-day dairies show that respondents were less accurate, and nonresponse was higher during the second day than during the first day (Arentze et al., 2001; Chatzitheochari et al., 2017). Alternatively, they may drop out altogether; nonresponse has been shown to be problematic in diary studies (e.g., Elevelt, Lugtig & Toepoel, 2019; Thompson et al., 2014; Van Ingen, Stoop, & Breedveld, 2008). Therefore, diaries are most often used for capturing only a couple of days of behavior (Schlich & Axhausen, 2003), although covering fewer days increases the chances that low-frequency behavior will not be captured (Gershuny, 2012).

Location data can be a valuable addition to time use research and diary studies in general. With these data, we get an insight not just into what people are doing but also where they are doing it (Plowman & Stevenson, 2012). This allows data to be understood in its context (Chen, 2011). When respondents, for example, fill out to be listening to music, the geographical location data can enable researchers to make a distinction between listening to music in the train or at home.

Many European TUSs follow the Harmonized European Time Use Survey (HETUS) guidelines (Eurostat, 2009). These guidelines recommend to code functional locations such as in transit, at home, working place or school, other people's home, or unspecified location. According to these guidelines, functional locations are not asked directly but determined by trained coders based on diary information. This is a very time-demanding, money-demanding and rater-dependent process. Alternatively, when respondents are asked to report their functional locations themselves, a lot of item missings are found (Chatzitheochari et al., 2017; ONS, 2006). Therefore, it would be useful to see whether we can automatically code functional locations, without interference of raters or the respondent himself or herself.

The use of smartphone-based diaries can, combined with sensor data, be an efficient and cost-effective solution to reach a large sample and follow them over a longer time without increasing respondent burden (Patel, Nowostawski, Thomson, Wilson, & Medlin, 2013; Raento et al., 2009). Researchers can passively record geotracking information by using the location services of the smartphone. Geotracking enables researchers to see exactly

where participants are and how they travel throughout the day (Bohte & Maat, 2009; Miller, 2012). Respondents' geolocations can be determined both by the GPS receiver incorporated in smartphones and by cellular towers and Wi-Fi networks. All methods result in a location that consists of a latitude and longitude measurement. Depending on the method being used, locations can be measured accurately with a precision of about five meters (GPS/Wi-Fi), although precision is much lower when cellular networks are used. Geotracking records locations at regular intervals and will therefore yield a very large and detailed data set of location coordinates of every respondent, from which we can potentially automatically derive functional locations.

Researchers have already started to investigate how to combine geographical location data and survey data in mobility research. The purpose of location data in mobility research is to identify movement or trips. Respondents are asked to write down the locations of the beginning and destination of the trip, along with the purpose, length, and distance of the trip and mode of transport. This purpose is different from recording locations of time use, but there is still a lot we can learn from mobility research. Traditionally, mobility research data were also collected through paper travel diaries in which respondents were asked to record their travel behavior for several days (Bohte & Maat, 2009). GPS-based data methods are potentially more accurate and less burdensome for respondents compared to traditional paper travel diary methods. The evaluation of a GPS-based mobility survey confirms that this method has a lot of potential (Bohte & Maat, 2009; Cottrill et al., 2013). More exact data about routes, trip length, and trip duration can be collected, as these are really hard for respondents to estimate precisely. Subsequently, these data can be enriched with a survey in which respondents are asked about the functional locations or origin and destination, and trip purposes.

Problems identified in studies from mobility research include the need for rule-based algorithms, GPS trackers' battery life, and geographical location data quality. Because of the large quantity of location data that can be collected, there is a need for procedures to manage, manipulate, and analyze the data. We need to set up rules and algorithms to automatically detect trips and determine trip mode and purpose (Stopher, Fitzgerald, & Zhang, 2008). Another issue is smartphone or GPS logger battery drain due to geotracking. Respondents need to charge their device (more) often, which they may forget or perceive as burdensome, inducing nonresponse (Bohte & Maat, 2009; Cottrill et al., 2013). Furthermore, the geographical location data can contain measurement error. GPS coverage varies, and GPS accuracy depends on satellite geometry, signal blockage, atmospheric conditions, and receiver design features/quality (in other words, the device itself). Geographical location data may be inaccurate as GPS trackers may, for example, lose signal indoors, in trains, or in areas with many tall buildings (Bohte & Maat, 2009; Zheng, Li, Chen, Xie, & Ma, 2008), leading to positioning errors (Song & Lee, 2015).

According to Stopher and Shen (2011), location trackers sometimes miss trips that are of short duration or distance or when the time between trips is short. The device then fails to locate the position before the trip has ended or the next trip has started. Next to these technical drawbacks, the problems can be caused by the respondent who can turn off the location tracker or forget to carry the location tracker during travel.

In this study, we assess whether locations measured by smartphones are sufficient to automatically record functional locations of respondents and see whether we can overcome some of the problems of geotracking identified in mobility studies. The use of location data next to time use day could add context to the time use codes (Chen, 2011). We show whether and how we can use geographic data to automatically record functional locations in time use research. In order to test this, we apply a set of simple rules to assign functional locations in different steps to the raw location data and link this to the time use data. When functional locations are coded to be the same in both data sources, we consider them to be correctly classified. We start by investigating the correspondence between the frequency distributions of functional locations assigned by the location data and by the time use diary data. Second, we investigate whether the functional locations are also correctly classified as we compare them at a fixed ten-minute interval level. Third, we investigate for what specific activities we find correspondence between the functional locations. Finally, we try to explain the variability by investigating for whom and when we can automatically record functional locations.

4.3 Method

4.3.1 Sample and Procedure

To answer the question whether we can automatically record functional locations from geotracking data, we use self-reported time use data and passively collected tracking data of young adults. We invited panel respondents of the ongoing Children of Immigrants Longitudinal Survey in the Netherlands (CILS4EU; Jaspers and Van Tubergen, 2015; Kalter et al., 2016) to participate in a smartphone survey study lasting 4 days. The CILS4EU panel study started in 2010 with a sample of 14-15-year-olds (third-year high school pupils in the Netherlands) that oversampled immigrant minority youth. The name of the survey that was used to invite respondents was Youth in Europe Study. In the Netherlands, respondents were followed up each year until Wave 5 for which fieldwork was completed in the period January to May 2015. 2,658 randomly sampled panel respondents, for whom there was a valid e-mail or postal address, were invited to our smartphone study. The unique login codes that were sent to the respondents corresponded to a randomly assigned set of four fieldwork days in the two-week period of September 28 to October 11, 2015. The four days were always at least two days apart. Invitation letters were sent, so that

respondents received them one week prior to the first fieldwork day, which would give them sufficient time to install the app and login for the first time. The app then retrieved the fieldwork days assigned to the respondents automatically from the server. All those who had not yet logged in for the first time on Monday, September 28, 2015, were sent a reminder e-mail. Anyone who had already missed the first-week fieldwork day was assigned two additional fieldwork days in a third week of data collection (October 12 to October 18, 2015).

We used a progressive remuneration scheme in which we sent respondents a 40-euro gift card if they completed the full four days of the TUS. Those who missed one day received a 20-euro gift card, and respondents who participated for two days got a 10-euro gift card. If respondents participated fewer than two days, they did not receive a gift card.
Of the 2,658 invited panel respondents, 50.4% participated in our smartphone diary study (N 1,339); 83.5% of the respondents had at least one observation on each of the four fieldwork days, 7.7% on three days, 4.2% on two days, and 4.7% on one day.

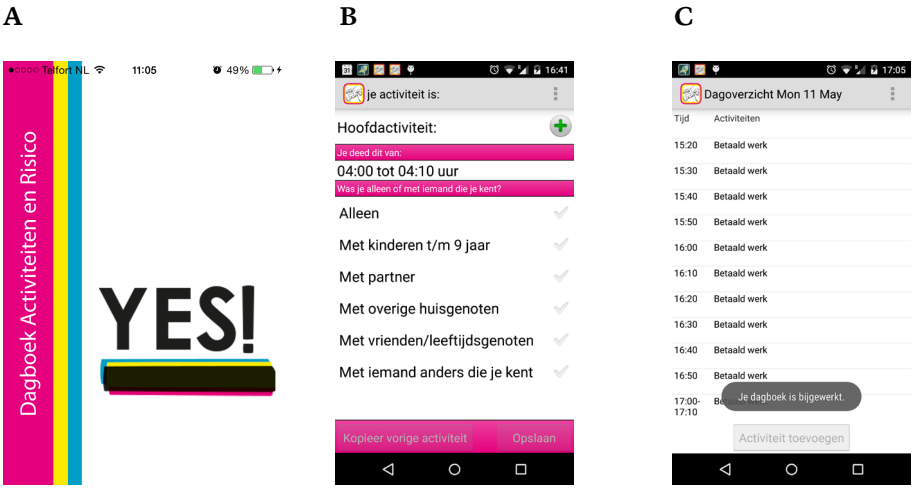


Figure 4.1 Screenshots of the time use survey app. (A) The login screen; (B) the screen where activities were reported; and (C) the day overview of a set of recorded activities.

4.3.2 The Smartphone Time Use and Victimization Survey App

In order to study in greater detail the situational factors that influence the risk of victimization, the Netherlands Institute for the Study of Crime and Law Enforcement (NSCR) developed a dedicated time use and victimization smartphone survey app Dagboek Activiteiten en Risico [Activities and Risk Diary]. The app conforms to the HETUS guidelines on harmonized European TUSs (Eurostat, 2009) in that it asks

respondents to report about their activities in ten-minute intervals, starting each response day at 4:00 a.m. It uses the overall HETUS activity categories (for more details on the app design and how it follows HETUS guidelines, see Sonck & Fernee, 2013; see Figure 4.1 for screenshots of the TUS app). Although the app could be installed by anyone, it could only be used with a unique login code that was sent to respondents in the invitation to participate.

The app was programmed by CentERdata, an independent research institute in the Netherlands, who also hosted the servers to which all responses were sent. Commissioned by NSCR, CentERdata extended the app with additional questions on victimization, witnessing crime, and substance use, additional answer categories for the questions on mode of transportation and other people present. These data are not used in this study.

Following the HETUS guidelines, the time use diary divided the day into 144 fixed ten-minute timeslots. The total number of observations available for analysis was 702,832 ten-minute intervals (timeslots), implying that the 1,339 respondents on average provided data for 525 ten-minute time intervals (91% of all possible timeslots).

Table 4.1 Average percentage of time spent on various activities per day.

	Dutch TUS 2016		**Our sample**	
	12 – 19 years	20 – 64 years	iOS users (18 – 24 years)	Android users (21- 24 years)
Personal Care	48.1	44.9	50.6	49.5
Employment	7.3	16.6	7.4	8.1
Study	13.3	1.85	10.9	11.0
Domestic work and care taking	4.05	8.93	3.43	3.1
Leisure	26.4	25.77	21.1	21.9
Volunteer work	0.48	1.55	0.3	0.3
Travel			6.4	6.1

Note. 63.4 percent (n = 849) of the respondents used an iOS-device, 36.6 percent (n = 490) of the respondents used an Android device.

4.3.3 Comparison of Respondent Activities with the Dutch TUS

By comparing the reported activities from our sample with data from the Dutch TUS 2016 (Roeters, 2017), we can estimate whether the smartphone diary app generated results comparable to that of the official Dutch TUS. The average age in our sample was 20.55 years old, ranging between 18 and 24. We compare our sample to two separate age categories of the Dutch TUS: 12–19 years, as the lifestyle of our sample is probably most comparable to this group (education and few responsibilities concerning domestic work and care) and 20–64 years, as the average age in our sample lies in this range.

The Dutch TUS divides the separate activity codes into six main categories: personal care, employment, study, domestic work and caretaking, leisure and volunteer work. Table 4.1 shows that the reported activities of our sample are quite similar to the activities of the 12–19-years-old group in the Dutch TUS 2016. The main differences lie in the time spent on domestic work and leisure, which is lower in our sample. This difference can be explained by the absence of the main category "traveling" in the Dutch TUS, where travel episodes were classified based on the trip purpose (e.g., traveling to a sport club would be classified as leisure in the Dutch TUS).

4.3.4 Time Use Location Codes

We assigned all 41 time use activity codes to one of five functional locations based on where the activities are expected to happen. The United Kingdom Office of National Statistics (2006) uses a crude distinction between activities at home versus those away from home. We subdivided the away category further into at work/school, in transit, and at some other location. We also included a category everywhere for activities that have no clear functional location and could thus occur in all locations (see Appendix, Table 4.14) for the classification scheme for all 41 HETUS activity codes. To our knowledge, ONS (2006) is the only organization using respondents' self-reports of their locations and time use following the HETUS guidelines.

Home included all activities related to personal care or the household, such as sleeping, washing, and dressing, or doing housework. Following reports of ONS (2006), media- or computer-related activities such as watching TV, reading, and using the computer were also assigned to the *home* category.

Work/school is expected to be the main daytime activity of most people in our sample. However, some respondents are doing vocational training at work, so we combined the two.

Other included all activities that we expected to happen at another location than the *home* or *work/school* locations. This category includes activities related to leisure time, going out, and doing volunteer work, for example, shopping, cultural visits, sports, religious activities, and helping others outside the family.

In transit included all recorded trips, traveling by own means/transport and traveling by public transport reported by respondents.

Everywhere included all remaining categories that could reasonably take place at any location. This category includes having a talk, using the telephone, registering time use, and listening to radio and music.

4.3.5 Location Tracking

At installation of the smartphone survey app, respondents were asked whether they agreed with sharing their geotracking information over the four-day period when they completed the time use diary; 94.4% of the participants gave consent to share GPS data (n = 1,264). The location tracking services of the smartphone platforms were then used to record time-stamped longitude/ latitude coordinates.

In total, 941,821 location tracking measurements were collected (*mean* per person = 726.6). The frequency of measurement differed between the two platforms. Android phones record GPS coordinates every 10 min, which in theory would sum to 576 GPS locations for the four days of data collection. However, the location tracking services did not always get a valid measurement immediately (e.g., in densely build areas or indoors), and if that occurred, the app retried to get a valid measurement of coordinates a few minutes later. This resulted in more measured coordinates. On average, 986.3 coordinates were available per Android user (472,422 in total). On iOS devices, the location services operated differently. The measurement of time-stamped coordinates only starts when the accelerometer on the smartphone detects large movement, after which GPS coordinates are recorded frequently until the movement stops. On average, 573.1 coordinates were collected per iOS user (469,388 in total).

4.3.6 Analysis of Correspondence Between Diary and Geotracking Locations

In order to link the geotracking data to the time use diary data, we split the geotracking data in the same ten-minute timeslots as used in the time use diary. When there were multiple coordinates measured within a particular timeslot, the first measurement within the timeslot was used.

In merging the time use data with the coded geotracking information, we only kept those timeslots for which we at least had information from the TUS. Missing geotracking data had a different meaning for iOS and Android users, due to the different ways of measuring. iOS devices only stored geotracking information when movement was detected by the accelerometer of the phone. This implies that no coordinates were recorded when a respondent was stationary. However, missing data could also mean that a respondent turned off his/her phone or GPS tracker. We imputed all stationary time GPS data from iOS devices though using the "na.lofc" function (Zeileis & Grothendieck, 2005) in R, replacing all missings with the most recent nonmissing value prior to it.

4.3.7 Functional Locations Derived from Geotracking Data

In order to derive functional locations from the time-stamped GPS coordinates, we first had to deal with small measurement errors in the geotracking data. This was done by rounding the longitude/ latitude coordinates to the first three decimals, which limits spatial precision to 111.32 m. These rounded data were subsequently used to derive four location categories for the GPS data.

All coordinates that were recorded when a movement of at least 100 meters was detected were coded as *in transit*.

Home was coded as the main location where respondents were for at least 10 timeslots (1 hr and 40 min) between 4:00 and 6:00 in the morning.

Work/school was coded as the main location where respondents were for at least 15 timeslots (2 hr and 30 min) between 10:00 and 15:00 in the morning/afternoon.

After assigning the *home* and *work/school* location this way, all timeslots that had the same coordinates at later time points were also coded as *home* and *work/school,* respectively. All other locations, where the respondent was not in transit, not at home, nor at school/work, were coded as *other.*

4.3.8 Analytical Method

Because our main objective is to investigate whether it is possible to automatically infer functional locations from geotracking data, the correspondence between the functional locations derived from the self-reported time use data with those derived from the geotracking data needed to be assessed. This was done by crosstabulating the codes from the two different sources and subsequently calculating the percentage that was coded to be the same. When timeslots were coded the same in both data sources, they were counted as

correctly classified.

In order to investigate whether the two data sources match, we calculated the total accuracy by dividing the number of correctly classified timeslots by the total number of timeslots assigned to *home, work/school, other,* and *in transit*. In this calculation, we excluded the category everywhere because it is impossible to assess correct classification for these activities as they can take place everywhere. In order to investigate which functional locations we can automatically derive, the *total percentage correct* is calculated by dividing the matching timeslots of a specific location by the total number of timeslots assigned that specific location by the TUS.

We subsequently examined which types of activities, classified according to the HETUS codes, were better classified than others. Finally, we estimated several multilevel logistic regression models in order to assess whether correct classification mainly varies within or between individuals. This allows us to assess when and for whom it is possible to automatically record functional locations from geotracking data.

All data cleaning and analyses were done in RStudio (R Core team, 2020).

4.4 Results

4.4.1 Linking Location and Time Use Data

We start by investigating whether the aggregated frequency distributions of the functional location codes are the same for the TUS and the location tracking. The frequency distributions of the time use and location tracking data shown in Table 4.2 are more similar for Android users than for iOS users. This is mostly caused by the large overrepresentation of the category *other* for iOS according to the location tracking, which covers almost half of the records of iOS users. The functional location category *other* is also overrepresented in the Android data, but much less so than in the iOS data. For Android users, the percentage *home* and *in transit* are almost the same for the TUS and location tracking. For both types of devices, the functional location code *work/school* is underrepresented; the percentage of work or school locations assigned based on the location tracking is only half of the percentage according to the TUS. This suggests that we succeed only half the time in automatically assigning a school or work location. There are multiple possible reasons why we fail to find high correspondence between the measures. Either the diary data are incorrect or—in our view, more likely—the GPS data cannot be easily assigned to a functional location. To further investigate this issue, we conduct several additional analyses later in this article. First, we investigate matches at the level of the timeslot to inspect in more detail how data are misclassified.

Table 4.2 Average percentage of time of the day spent at various functional locations, as assigned based on the time use survey and location tracking.

	iOS Users		Android Users	
	Time Use Survey	Location Tracking	Time Use Survey	Location Tracking
In transit	6.7	17.0	6.5	9.1
Other	7.8	46.9	7.3	28.0
Home	56.6	26.7	56.6	53.7
School / Work	16.2	9.4	17.2	9.3
Everywhere	12.7		12.4	

4.4.2 Linking Location and Time Use Data at the Timeslot Level

In the first step, we assigned functional locations to timeslots per day. This implies that what is referred to as the respondent's *home* and *work/school* locations can vary from day to day (e.g., someone may have different work locations). The total percentage of correctly classified timeslots for iOS users is 33.3% and for Android users 58.5%.

Tables 4.3 and 4.4 show the results and correspondence at the timeslot level in more detail. What stands out of this table is that there is much more correspondence between time use diary and GPS locations for Android phones than for iOS phones. For Android data, the *home* location is correctly estimated for 70.5% of time use reports, while for iOS, this is 35.6%. *Work/school* are not consistently categorized correctly (21.3% and 32%). Especially for iOS users, respondents are far too often at another location according to GPS when they are at home according to time use data. We fail to automatically record a home location for these participants. Next, we investigate several potential reasons why our method does not work.

First, there can be coding errors in the geographical location data. For example, respondents are not at home or work on a particular day, so we cannot assign a home or work location successfully. We therefore repeat our analysis, now coding over days, by using the geotracking data of all recorded days. We find that this does not solve the problem of low correspondence of measures though. The total accuracy for iOS is 24.7%, and the total accuracy for Android users is 52.8% (see Table 4.8 and 4.9 in the Appendix) for the complete results of these analyses.

Second, there can be coding errors in the time use data. For example, respondents fill out activities they do not actually do. We cannot control for that here and come back to this in the discussion.

Third, the two data sources are out of sync. For example, respondents miscode their time use activities due to recall error; an activity is recorded at a particular timeslot, when in reality it had happened a timeslot earlier or later. Allowing for these small temporal inconsistencies after matching the two data sources improves the coding, implying that respondents sometimes misclassify the timing of their activities. A more liberal coding by definition results in more matches and higher correspondence throughout. The total accuracy for iOS users increases to 36.0%, and the total accuracy for Android users increases to 61.3% (see Tables 4.10 and 4.11 in the Appendix for the complete results of these analyses).

A fourth possible reason is that the location data do not provide enough information in isolation. So far, we have assumed that the TUS data are correct and used those to better code the functional locations of the GPS measurements. Incorporating respondent information in our model, by using the locations where respondents filled out they were sleeping, working, and studying, to classify other locations at different times and days largely improves the coding of *work* and *school*. The total accuracy also increases but mostly due to the improvement in the coding of *work* and *school*. The total accuracy for iOS users increases to 44.9% and for Android users to 64.5% (see Tables 4.12 and 4.13 in the Appendix for the complete results of these analyses). Finally, we chose to use the results as reported in Tables 4.3 and 4.4 for the remainder of the article.

These results are not the most accurate, but the increase in accuracy by more liberal coding is not very large. The remainder of the article focuses on investigating for what specific HETUS codes, the GPS locations match.

Table 4.3 Matching per participant, per day, iOS users.

		Codes based on TUS				
		In transit	Everywhere	Other	Home	Work
Based on GPS	In transit	**6196**	9849	6918	28132	15165
	Other	14069	25975	**15146**	101206	25971
	Home	2922	9045	5010	**78195**	8440
	Work	2912	4660	3098	12337	**13420**
Total percentage correct		23.7		50.2	35.6	21.3

Table 4.4 Matching per participant, per day, Android Users.

		Codes based on TUS				
		In transit	Everywhere	Other	Home	Work
Based on GPS	In transit	**6443**	2964	2670	4949	4064
	Other	5030	**9673**	**7205**	29043	14169
	Home	3049	14319	5891	**92859**	8947
	Work	682	2030	1206	4936	**12781**
Total percentage correct		42.4		42.5	70.5	32.0

4.4.3 Time Use Categories

Overall, we detect no large differences between activities that fall within the same functional location category. We see that we can assign functional locations generally better for Android users than for iOS users. The percentage of timeslots correctly classified is relatively low for iOS users; even for an activity like sleeping. For Android users, we are able to assign *home* codes quite well but not *work* and *school* locations (see Appendix, Table 4.14), for the complete results per HETUS category.

4.4.4 Multilevel Models: Exploring Variability

To explore the variability within the (in)correct classifications, we ran several multilevel models. We predict whether there is a match (0=no match, 1=match) between the functional location assigned by the location data and the functional location based on the time use data. As we have many measurements per person (one for each timeslot), and these measurements may not be independent, we treat time as Level 1 and the person/participant as Level 2 in our multilevel model. In addition, we calculate the intraclass correlations (ICCs) to investigate the proportion of within (time) and between (person) variance (Level 1 = time, Level 2 = person; see Tables 4.5 and 4.6 for the results of the multilevel models per phone type).

In Model 1, we add time and type of day as Fixed Level 1 effects to our model. Daytime divides the day into night (0:00–6:00), morning (6:00–12:00), afternoon (12:00–18:00), and evening (18:00–0:00). We made three dummies (morning, afternoon, and evening) using night as reference category. Day type is a dummy representing the difference between weekdays (=0) and weekends (=1). In Model 2, we added the number of times

the respondent filled out the diary as a Fixed Level 1 and Fixed Level 2 effect. To use the number of times the diary is filled out as a Level 1 effect, we calculated *the number of times the diary was filled out* at a specific day of participation and centered that value within persons. As Level 2 predictor, we used the grand mean centered average number of times the diary was filled out per respondent per day.

We started with the intercept-only model and calculated the ICC (Hoffman, 2015; Snijders & Bosker, 1999). The ICC is a standardized way of expressing how much dependency is due to person mean differences (Hoffman, 2015). We calculated the ICC as *(τU02)/(τU02 + 3.29)* (Hoffman, 2015; Snijders & Bosker, 1999). In our case, 11% (Android) and 15% (iOS) of the original outcome variation can be explained by between-person mean differences over time. So the largest proportion of outcome variation is due to within-person differences.

Results of Model 1 show that for both phone types, more matching occurs at night than in the afternoon and evening. For the Android users also more matching occurs at night compared to the morning, whereas for the iOS users, more matching occurs in the morning. In addition, more matching occurs on weekend days than on weekdays. Results of Model 2 show that the number of times a respondent filled out the diary does not predict matching success. When looking at the model fit indices, it can be observed that the decrease in AIC is only small, indicating that adding the number of times the diary is filled out as a predictor does improve the model only slightly.

After running these three models for iOS and Android separately, we merged the data of both groups together. See Table 4.7 for the results of the multilevel models for both groups together. The first intercept-only model shows that the ICC is higher than in the separate models. This larger variance between persons is mostly explained by phone type, as can be observed in our second model; iOS users have fewer matches than Android users. When we add phone type as Level 2 predictor, the ICC decreases from 18.6% to 14.4%. The same effects of type and time of day are observed as in the models per device type; timeslots earlier in the day (night and morning) and in the weekends are more often classified correctly. Finally, we added the number of times the respondent filled out the diary that day as fixed Level 1 and Level 2 effect to our model. There is a negative effect, but like previous models, the small decrease in AIC indicates that adding the number of times the diary is filled out as predictor improves the model only slightly. Chi square tests revealed that for all data sets, Model 1 had a better model fit than the intercept-only model, and Model 2 had a better model fit compared to Model 1.

Table 4.5 Multilevel Logistic Regression Models Predicting Match between GPS and TUS data, for Android Users.

	Model 1		Model 2		Model 3	
	B	**SE**	**B**	**SE**	**B**	**SE**
Fixed Coefficients						
Intercept	0.04*	0.03	0.68*	0.04	0.71*	0.04
Daytime						
Morning			-0.24*	0.01	-0.24*	0.01
Afternoon			-1.40*	0.01	-1.40*	0.01
Evening			-1.39*	0.01	-1.39*	0.01
Daytype						
Weekend			0.08*	0.01	0.07*	0.01
Times filled out						
Difference from average					0.007	0.01
Average per person					-0.003	0.01
Random Effect						
Intraclass correlation	.106		.123		.123	
AIC	3016132.5		285170.1		285101.7	

Note. TUS = time use survey; GPS = global positioning system; AIC = Akaike information criterion.*$p <$ .001.

Table 4.6 Multilevel Logistic Regression Models Predicting Match between GPS and TUS data, for iOS Users.

	Model 1		Model 2		Model 3	
	B	**SE**	**B**	**SE**	**B**	**SE**
Fixed Coefficients						
Intercept	-0.99*	0.03	-1.01*	0.03	-1.01*	0.03
Daytime						
Morning			0.13*	0.01	0.13*	0.01
Afternoon			-0.78*	0.01	-0.78*	0.01
Evening			-1.05*	0.01	-1.05*	0.01
Daytype						
Weekend			0.30*	0.01	0.29*	0.01
Times filled out						
Difference from average					0.01	0.01
Average per person					-0.004	0.01
Random Effect						
Intraclass correlation	.148		.157		.157	
AIC	440900.9		422590.7		422564.1	

Note. TUS = time use survey; GPS = global positioning system; AIC = Akaike information criterion.*$p <$ .001.

Table 4.7 Multilevel Logistic Regression Models Predicting Match between GPS and TUS data, for Android and iOS Users.

	Model 1		Model 2		Model 3	
	B	**SE**	**B**	**SE**	**B**	**SE**
Fixed Coefficients						
Intercept	-0.60*	0.02	-0.78*	0.02	-0.78*	0.03
Phonetype						
Android			1.10*	0.04	1.12*	0.05
Daytime						
Morning			-0.01	0.01	-0.01	0.01
Afternoon			-1.04*	0.01	-1.04*	0.01
Evening			-1.18*	0.01	-1.18*	0.01
Daytype						
Weekend			0.21*	0.01	0.20*	0.01
Times filled out						
Difference from average					0.005	0.004
Average per person					-0.003	0.004
Random Effect						
Intraclass correlation	.186		.144		.144	
AIC	747546.8		709386.2		709301.8	

Note. TUS = time use survey; GPS = global positioning system; AIC = Akaike information criterion.*$p <$ .001.

4.5 Discussion

In this article, we investigated whether the passive collection of location tracking data is sufficient to automatically establish functional locations in time use research. This study contributes to developing a methodological framework for integrating sensor and survey data, in particular with respect to large-sample time use research. We know little about how to effectively analyze and use mobile phone sensor data in large-sample studies. This study provides a step forward toward the integration of multiple data sources.

Our results show that by integrating time use and geographic location, we can to some extent derive functional locations automatically, at least for Android users. For iOS users, the results of the automated coding of functional locations were rather disappointing. An important limitation for automatically deriving functional locations in this TUS is that measurement errors occur in the location tracking for both phone types (Android and iOS). Therefore, in order to adequately record functional locations, active input from respondents is still needed.

The automatic coding of home location performs well for Android users, so it seems that we are reasonably successful in determining a person's home location. The automatic coding of work and school does not perform well, but the automatic coding of these locations improves when we incorporate information from the respondent. Four days may not have been sufficient to improve our model; however, future research could investigate the effects of asking respondents for GPS locations of their main locations (home, work, and school) or collecting data for more days on automatic assignment. Zenk, Matthews, Kraft, and Jones (2018) recommend to measure location data for at least 14–28 days to measure respondent's activity spaces. Or, alternatively, we need respondent information to better assign these locations.

Overall, measurement error in location tracking, also called positioning error, proved to be a major issue. Location trackers can suffer from positioning errors (Song & Lee, 2015), especially indoors. Even though we rounded our GPS data to 112 m, we still found small position changes at times when one would expect most respondents to remain stationary (e.g., at night and at work/ school). An indication of measurement error in our data is the overrepresentation of the functional location category *other*, as this indicates that too many locations were not assigned to a home, work, or school location. This might be due to a larger variability in functional locations than anticipated. However, it is more likely due to measurement error in the location tracking. When positioning errors occur, the measured GPS coordinates for a specific timeslot are (slightly) incorrect. In that case, the GPS coordinates of that timeslot cannot be matched to the GPS coordinates of the home or work location of the respondent, and the timeslot is incorrectly assigned the functional

location *other*. We advise future research to focus on detecting, eliminating, and correcting positioning errors to improve data accuracy. Some studies have recorded the precision of individual measurements and have used these in modeling the location data. This seems to be a promising way to deal with possible errors in future research.

The variability in matching success is mainly intrapersonal. Results of the multilevel model in which we predict for which timeslots we can automatically derive functional locations show that only 10–19% of the variance is between persons. Most of the variance is thus within persons, making it difficult to predict which time–location data is difficult to match. Our multilevel model shows that time of day is a particularly important predictor, which may indicate that matching success is related to complexity. The night and morning may be relatively easy to predict as these follow the same daily cyclical structure of sleeping, waking up, and going to work or school. Respondents may change locations relatively little, compared to the afternoon and evening. Respondents have more leisure time later during the day (Cloïn et al., 2013) and therefore more alternative options for locations, which complicates the automatic derivation of functional locations. This argument seems to speak against our finding that functional locations are easier to predict on weekends because weekends contain more leisure time. However, weekends are also characterized by less variation in activities, which may decrease overall complexity (Sonck & Fernee, 2013). A larger complexity in activities could also lead to an increase in survey fatigue and satisficing behavior (Krosnick, 1991). Satisficing respondents do not try to recall all that is relevant, but just enough to provide a reasonable answer, leading to a deterioration in time use data quality. Incorrect matches might thus also be partly due to incorrect time use reports. Future research could shed more light on the effect of complexity and on the validity of the various explanations of the large intrapersonal variance.

Phone type strongly affects the accuracy of geolocation measurement. This is demonstrated by the large differences in matching percentages between iOS and Android users and also by the large predictive value when we include this variable in our multilevel model. This phone type effect is likely due to the difference in location geotracking strategies of both platforms; Android phones record GPS coordinates every 10 min, whereas iOS phones record GPS coordinates when movement is detected. iOS phones may have suffered from a "cold/warm start" problem (Stopher, Fitzgerald, & Zhang, 2008): some amount of travel has already been completed before the first new positions are being recorded and stored. In that case, our app may have recorded the wrong geolocation and have failed to detect the exact stationary location. For future research, we advise to combine the two geotracking methods. Energy-wise, it is very inefficient (and also complicates data storage and analysis) to record geolocations constantly at a fixed rate, although an interval of ten minutes between measurements should be reasonable.

Additionally, when movement is detected, the frequency of recordings could be increased to capture trips in greater detail. More geolocations are necessary if exact routes need to be recorded at finer granularity. For researchers interested in mobility, iOS data may currently be more useful.

Many decisions were made to align the sensor data with the survey data. This may have affected our results. For example, rounding GPS data and choosing a GPS measurement per timeslot may affect whether or not we are able to derive functional locations correctly. Without rounding, or with rounding on more than three decimals, small positioning errors decrease the successful classification of home and work locations. But rounding may also have increased the number of in transit codes. Rounding to three decimal places created 112 meters of precision, yet travel was assumed any time two data points were more than 100 meters apart. If two locations are both slightly off, but one is rounded up and one is rounded down, it may incorrectly seem like the respondent is in transit. This may have increased the incorrect number of *in transit* codes. Alternatively, the travel distances could be calculated before rounding and before coding the home and work location. Another important decision we made was using only the first GPS measurement per timeslot if more measurements were available. Android users had only one measurement available and we wanted to treat the iOS users similarly, but one could also compute the center point or check all GPS locations when more are available. A finer granularity of location tracking may increase the correspondence between the two data sources. However, more frequent GPS locations would make it more difficult to match timeslot diary data to GPS-derived locations.

Unfortunately, we only had proxy information about respondents' locations available. We could not let respondents validate our functional location codes, so we used information recorded in the TUS instead. Based on previous research (ONS, 2006) and common sense, we assigned the activity codes to one functional location, but these activities might occur at other locations as well. When we investigate the percentage of matching per activity, however, there are no activity categories that are assigned to another functional location category more often, supporting our assignment strategy. Furthermore, like every diary study, recall bias may occur and respondents may not actually have performed the activities they reported. When we correct for that, by allowing small inconsistencies, the accuracy only increases a little. Future research could ask respondents to validate their location or keep track of their locations. However, this is a very expensive and burdensome process especially in a large-sample study.

A last limitation of our study is the specific type of sample that was very willing to share GPS data (94.4%). This will probably not be the case in general population studies. This type of study may therefore be less feasible in random samples of the general population.

Finally, to answer our main research question, can we reliably automatically record functional locations by combining activity recordings with geotracking? Due to measurement errors in geo-tracking, this is possible only to some extent (notably places of residence can be established), and respondents' self-reported information is still necessary to measure other functional locations. Future research could explore different methods of filtering positioning errors and recording GPS coordinates, and variations in recording frequency length, to investigate its effect on data quality. Location tracking data may be easier to process in the future and more reliable to use to better fulfill the potential to enrich smartphone survey data collection.

4.6 Appendix

Table 4.8 Matching over days, iOS users.

	Codes based on TUS					Percentage GPS
	In transit	**Everywhere**	**Other**	**Home**	**Work**	
Based on GPS						
In transit	**6731**	10704	7501	31934	16328	18.8
Other	15461	26856	**16159**	118470	33694	54.2
Home	2358	8190	4413	**53528**	5737	19.1
Work	1549	3779	2099	15938	**7237**	7.9
Percentage TUS	6.7	12.7	7.8	56.6	16.2	
Total percentage correct	25.8		53.6	24.4	11.5	

Table 4.9 Matching over days, Android users.

	Codes based on TUS					Percentage GPS
	In transit	**Everywhere**	**Other**	**Home**	**Work**	
Based on GPS						
In transit	**6727**	3437	3224	5695	5132	10.4
Other	5592	**10628**	7591	35548	19793	34.0
Home	2403	13418	5587	**85327**	7065	48.9
Work	482	1503	570	5217	**7971**	6.8
Percentage TUS	6.5	12.4	7.3	56.6	17.2	
Total percentage correct	44.2		44.7	64.8	20.0	

Table 4.10 Match and allow small inconsistencies, iOS users.

	Codes based on TUS					Percentage GPS
	In transit	**Everywhere**	**Other**	**Home**	**Work**	
Based on GPS						
In transit	**9055**	8942	6316	27429	14518	17.1
Other	12755	25081	**18371**	100332	25828	46.9
Home	2469	6556	4608	**81755**	8224	26.7
Work	2458	4041	3046	12191	**14691**	9.4
Percentage TUS	6.7	12.7	7.8	56.6	16.2	
Total percentage correct	33.9		56.8	36.9	23.2	

Note. We corrected for temporal inconsistencies by also counting a timeslot as correctly classified when the same activity code was recorded a timeslot earlier or later than reported by the respondent. When respondents report an activity one timeslot earlier or later than it is observed in the location data we replace codes based on the TUS of that specific timeslot with the correct location code.

Table 4.11 Match and allow small inconsistencies, Android users.

	Codes based on TUS					Percentage GPS
	In transit	**Everywhere**	**Other**	**Home**	**Work**	
Based on GPS						
In transit	**8251**	2431	2283	4515	3610	9.1
Other	4602	**9423**	8230	28750	14115	29.0
Home	2442	10473	5300	**98258**	8592	53.7
Work	520	1706	1183	4875	**13351**	9.3
Percentage TUS	6.5	12.4	7.3	56.6	17.2	
Total percentage correct	52.2		48.4	72.0	33.7	

Note. We corrected for temporal inconsistencies by also counting a timeslot as correctly classified when the same activity code was recorded a timeslot earlier or later than reported by the respondent. When respondents report an activity one timeslot earlier or later than it is observed in the location data we replace codes based on the TUS of that specific timeslot with the correct location code.

Table 4.12 Use information from the TUS for matching, iOS users.

	Codes based on TUS					Percentage GPS	
	In transit	**Everywhere**	**Other**	**School**	**Home**	**Work**	
Based on GPS							
In transit	**6731**	10704	7501	9344	31934	6984	18.8
Other	12429	22362	**14199**	9259	71130	7883	35.3
School	1665	1906	812	**10185**	3906	288	4.8
Home	4512	13470	7302	5154	**110576**	3359	37.1
Work	762	1087	358	106	2324	**10434**	3.9
Percentage TUS	6,7	12.7	7.8	8.8	56.6	7.4	
Total percentage correct	25.8		47.1	29.9	50.3	36.0	

Note. We assigned the code *home* to the main place where respondents had indicated in the time use survey they were sleeping. The main location where respondents had indicated they were working was assigned the code work, and the main location where respondents were *studying* was assigned the code *school*.

Table 4.13 Use information from the TUS for matching, Android users.

	Codes based on TUS					Percentage GPS
	In transit	**Everywhere**	**Other**	**School**	**Home**	**Work**
Based on GPS						
In transit	**6727**	3437	3224	2713	5695	2419	10.4
Other	4788	9285	**7242**	4870	24746	3084	23.2
School	745	1177	349	**8637**	2295	82	5.7
Home	2692	14622	5976	4843	**98302**	2615	55.4
Work	252	465	181	7	749	**10691**	5.3
Percentage TUS	6.5	12.4	7.3	9.0	56.6	8.1	
Total percentage correct	44.3		42.7	41.0	74.6	56.6	

Note. We assigned the code *home* to the main place where respondents had indicated in the time use survey they were sleeping. The main location where respondents had indicated they were working was assigned the code *work*, and the main location where respondents were *studying* was assigned the code *school*.

Table 4.14 Matching percentages per activity code.

Activity	TUS code.	iOS users.	Android users.
sleeping	Home	41.26	74.18
eat and drinking at home, work or school	Everywhere		
going out for eating and drinking	Other	60.47	66.99
personal or medical care	Home	19.79	65.54
employment	Work/School	21.71	37.29
school, university	Work/School	20.90	27.22
study, course as a hobby	Other		
cooking/food preparation	Home	19.79	62.78
household upkeep, cleaning	Home	21.23	68.11
gardening and taking care of pets	Home	22.83	47.65
DIY construction and repairs	Home	8.28	51.48
shopping/groceries	Other	47.73	40.27
services	Other	44.05	49.24
administration/paperwork	Home	26.59	69.33
caring and supervising children (of own family)	Everywhere		
helping other adults within own family	Everywhere		
voluntary work	Other	49.40	71.43
helping others, outside the family	Other	44.69	51.42
religious and ceremonial activities	Other	35.18	56.56
visits/having visitors, parties	Everywhere		
having a talk	Everywhere		
using the telephone	Everywhere		
going out, cultural visits	Other	66.56	62.10
library	Other	37.50	27.27
vising sports/competitions	Other	51.42	47.48
trips	In transit.	29.12	25.11
resting	Home	18.76	56.90
sports	Other	48.22	43.82
hobby	Other	50.22	34.06
gathering information and news via the internet	Everywhere		
online banking and online shopping	Home	18.97	56.35
communicating through the internet (online)	Everywhere		
other pc/internet offline	Home	30.16	64.40
playing games	Home	10.68	57.46
computer games	Home	23.95	58.38
reading	Home	24.08	70.49
watching television	Home	19.58	59.30
listening to radio and music	Everywhere		
travelling by own means/transport	In Transit	23.45	43.49
traveling by public transport	In Transit	*DOES NOT OCCUR*	
registering time use by smartphone	Everywhere		

Chapter 5

Squats in Surveys: Investigating the Feasibility of, Compliance with and Respondents' Performance on Fitness Tasks in Self-Administered Smartphone Surveys using Acceleration Data.

Elevelt. A., Höhne, J.K., Blom, A.G. (2020). *Submitted to Frontiers in Public Health.*

Author contributions: AE and JKH contributed to concept and design of the study. JH organized the data collection and database. AE performed the statistical analyses. AE wrote the first draft of the manuscript. JKH and AB critically reviewed the paper and wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Abstract

Digital health data that accompany data from traditional surveys are becoming increasingly important in health-related research. For instance, smartphones have many built-in sensors, such as accelerometers that measure acceleration so that they offer many new research possibilities. Such acceleration data can be used as a more objective supplement to health and physical fitness measures (or survey questions). In this study, we therefore investigate respondents' compliance with and performance on fitness tasks in self-administered smartphone surveys. For this purpose, we use data from a cross-sectional study as well as a lab study in which we asked respondents to do squats (knee bends). We also employed a variety of questions on respondents' health and fitness level and additionally collected high-frequency acceleration data. Our results reveal that observed compliance was higher than hypothetical compliance. Respondents gave mainly health-related reasons for non-compliance. Respondents' health status positively affected compliance propensities. Finally, the results show that acceleration data of smartphones can be used to validate the compliance with and performance on fitness tasks. These findings indicate that asking respondents to conduct fitness tasks in self-administered smartphone surveys is a feasible endeavor for collecting more objective data on physical fitness levels.

5.1 Introduction and Background

People's physical fitness level is crucial information in medicine and health-related research (Althoff et al., 2017; Bauman et al., 2009). When it comes to measuring physical fitness, most researchers rely on self-report questions employed in surveys (see International Physical Activity Questionnaire [IPAQ], 36-item Short Form Health Survey [SF-36], or LASA Physical Activity Questionnaire [LAPAQ]). For instance, the Health and Retirement Study (HRS, 2006) asks respondents the following question: "*Would you say your health is excellent, very good, good, fair, or poor?*" Such self-report questions are subject to respondents' own interpretation and evaluation of their physical fitness (Börsch-Supan et al., 2013; Kapteyn et al., 2018). In addition, Prince et al. (2008) suggest that self-report questions on physical fitness are prone to systematic measurement errors caused by social desirability (e.g., resulting in overreporting) or inaccurate recall (e.g., resulting in over- or underreporting). These methodological problems associated with subjective physical fitness measures in surveys exhibit the potential importance of more objective measures.

Replacing self-report questions with more objective measures on respondents' physical fitness level may decrease systematic measurement errors. Therefore, large-scale national and international health-related surveys, such as the Health and Retirement Study (HRS), the Survey of Health, Ageing and Retirement in Europe (SHARE), and the English Longitudinal Study on Ageing (ELSA), have regularly employed additional tasks to objectively measure respondents' physical fitness. In a pilot in 2006, the HRS, for instance, added several physical fitness tasks, such as a balance test (i.e., asking respondents to stand for ten seconds at a fixed point without stepping away from it) and a walking test (i.e., asking respondents to walk about two meters in a straight line), to its core survey modules. These tasks were overseen by and conducted with an interviewer present during the interview. Sakshaug, Couper, and Ofstedal (2010) reported that about 93% of the eligible HRS respondents complied with these fitness tasks. This high compliance rate might be due to the interviewer-administered survey setting. The presence of an interviewer may encourage respondents to participate in fitness tasks; a luxury not available in self-administered survey settings, such as web survey settings (Christensen, Elkholm, Glümer, & Juel, 2014).

Recently, many major interviewer-administered surveys, including major health-related surveys, switch to or experiment with self-administered web survey settings to be more cost and time efficient. For instance, since 2003 the HRS has assigned sub-samples of their respondents to participate in self-administered web surveys in an attempt to extend their ways of data collection.

This trend towards web survey settings opens novel ways to collect additional data that complement survey responses (Miller, 2012). This especially applies to mobile web surveys that are completed with mobile devices, such as smartphones (Dufau et al., 2011; Elevelt, Bernasco, Lugtig, Ruiter, & Toepoel, 2019; Elhoushi, Georgy, Noureldin, & Korenberg, 2017; Harari et al., 2016; Miller, 2012; Raento, Oulasvirta, & Eagle, 2009; Toepoel & Lugtig, 2015). Smartphone use in web surveys is rapidly increasing (Gummer, Quoß, & Roßmann, 2019; Revilla, Toninelli, Ochoa, & Loewe, 2016). From a measurement perspective, smartphones are attractive because they contain a variety of built-in sensors, such as accelerometers that measure acceleration, which is defined as the rate of change of velocity of an object over time. Acceleration data provide information about respondents' physiological states, such as movements, allowing researchers to infer respondents' completion conditions in surveys.

There is an increasing number of studies evaluating the usefulness and usability of acceleration data in smartphone surveys (Höhne & Schlosser, 2019; Höhne, Revilla, & Schlosser, 2020; Khan, Lee, Lee, & Kim, 2010). For instance, Höhne et al. (2020) investigated respondents' compliance with simple motion tasks, such as standing at a fixed point (as in a balance test) and walking around (as in a walking test), in a self-administered smartphone survey using acceleration data. The authors found compliance rates of about 90%, which correspond to the compliance rate of the interviewer-administered HRS 2006 pilot (see Sakshaug et al., 2010). In addition, the acceleration data of smartphones provided supporting evidence for respondents' compliance with the motion tasks.

The results from Höhne et al. (2020) indicate the general feasibility of fitness tasks in self-administered smartphone surveys to collect more objective measures of respondents' physical fitness. They also indicate that acceleration data of smartphones can be used to validate respondents' compliance with fitness tasks without requiring the presence of interviewers that oversee their completion. However, the small body of research on the compliance with fitness tasks, coupled with the limited number of fitness tasks tested so far, merits further investigation of the feasibility of fitness tasks in self-administered smartphone surveys.

In the present study, we go beyond existing studies and investigate respondents' compliance with doing squats (knee bends) for one minute. For this purpose, we conducted self-administered smartphone surveys in a field and a lab setting and collected high-frequency acceleration data of respondents' smartphones. Since the collection of the acceleration data occurs passively (in the background) there is no additional burden for respondents other than doing the squats and holding the smartphone during this task.

In what follows, we describe the research questions, the study design and passive data

collection, the task instructions and survey questions used, the underlying samples (cross-sectional study and lab study), and the analytical strategies. We then present the results of the study. Finally, we discuss practical implications associated with the feasibility of fitness tasks in self-administered smartphone surveys and address future research perspectives.

5.1.1 Research Questions

We start by making a distinction between hypothetical and observed compliance. While hypothetical compliance refers to respondents' general disposition to participate in a task, observed compliance, in contrast, refers to respondents' actual participation in a task. Empirical findings indicate that respondents' hypothetical compliance tends to be higher than their observed compliance with a task (Struminskaya et al., 2020a; Struminskaya et al., 2020b). Following this relation between hypothetical and observed compliance, we address the following research question: *Do hypothetical and observed compliance rates with fitness tasks in a self-administered smartphone survey differ from each other (RQ1)?*

Further, it is important to explore the reasons for non-compliance as these provide insights into respondents' decision process. Understanding respondents' reasons for non-compliance can help overcoming those reasons or encouraging respondents to comply in future studies. For instance, Höhne et al. (2020) investigated the reasons for non-compliance with simple motion tasks and found that respondents mainly reported issues related to health, surroundings, and situation. Thus, we address the following research question: *What are possible reasons for non-compliance with fitness tasks in a self-administered smartphone survey (RQ2)?*

Since unequal compliance propensities across key respondent groups may bias the sample it is important to investigate differences between respondents who comply and those who do not (Jäckle, Burton, Couper, & Lessof, 2017; Keusch, Struminskaya, Antoun, Couper, & Kreuter, 2019; Pinter, 2015; Revilla et al., 2016; Wenz, Jäckle, & Couper, 2017). In the HRS sample, for instance, respondents who complied with the additional fitness tasks (i.e., balance and walking tests) were more likely to be higher educated and had better self-reported health ratings (Sakshaug et al., 2010). Therefore, we address the following research question: *What respondent characteristics affect compliance with fitness tasks in a self-administered smartphone survey (RQ3)?*

In the HRS 2006 pilot, an interviewer has overseen the balance and walking tests to monitor respondents' compliance. As demonstrated by Höhne et al. (2020), however, such simple fitness tasks are also feasible in self-administered smartphone surveys. The authors argue that acceleration data of respondents' smartphones can potentially be used to monitor and validate respondents' compliance with fitness tasks without interviewers.

Accordingly, we address the following research question: *Can acceleration data be used to validate compliance with fitness tasks in a self-administered smartphone survey (RQ4)?*

The interviewer presence in the HRS 2006 pilot was not only important to the monitoring of respondents' compliance with the fitness tasks, but also to the monitoring of respondents' actual performance on the tasks. For instance, do respondents accurately perform the requested tasks or do they take shortcuts introducing measurement errors? Rowlands et al. (2018) have shown that acceleration data metrics from GENEActive accelerometers can be used as a complementary description of people's activity profile associated with fitness tasks and physical functions. The authors argue that acceleration data from smartphones are a useful source to evaluate respondents' performance on fitness tasks. Thus, we address a final research question: *Can acceleration data be used to validate respondents' performance (i.e. number of squats) on fitness tasks in a lab study (RQ5)?*

5.2 Method

5.2.1 Data Sources and Study Designs

In this study, we use two different data sources: Data from a cross-sectional study *(data source 1)* and data from a lab study *(data source 2)*. Both data sources contain high-frequency acceleration data collected from respondents' smartphones through the open-source JavaScript-based tool "SurveyMotion (SMotion)" developed by Höhne et al. (2020). SMotion collects the total acceleration (TA) of mobile devices, such as smartphones, on a survey page or question level, which is defined as follows:

$$\text{TA (Total Acceleration)} = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

Equation 1. Determining Total Acceleration (TA)

Note. Accelerations (a) along the x-, y-, and z-axis are defined as ax, ay, and az, respectively. The International System unit for acceleration is meter per second squared (m/s^2).

In this study, we calculated the average total acceleration for each respondent on the survey page on which respondents were required to do the squats. These average total acceleration values were based on the raw total acceleration data without checking for exceptionally low or high values because these values reflect specific characteristics of different motion levels that need to be preserved.

In general, the total acceleration of smartphones can be measured with and without gravity depending on the type of built-in accelerometer. Some old and/or low-budget devices are not equipped with accelerometers that contain a magnetometer (i.e., a sensor

for measuring gravity) that allows for the measurement of pure total acceleration without gravity. In these cases, only the total acceleration with gravity can be measured. We conducted all analyses using total acceleration data with gravity to keep the dataset as large as possible (Höhne et al., 2020).

The sampling rate of the total acceleration primarily depends on the device and/or on frequency restrictions set in the JavaScript code. In this study, the total acceleration of smartphones was measured without any frequency restrictions set in the JavaScript code to register it as precisely as possible. On average, the total acceleration was measured every 19 milliseconds.

In addition, we collected several types of paradata, such as response times, by using the open-source JavaScript-based tool "Embedded Client Side Paradata (ECSP)" (Schlosser & Höhne, 2018). Prior informed consent for the collection of total acceleration data and paradata was obtained by the survey company as part of panelists' registration process (cross-sectional study; *data source 1*). We also obtained informed consent in the lab study *(data source 2)*.

The dataset of the cross-sectional study *(data source 1)* serves for investigating respondents' (hypothetical and observed) compliance, reasons for non-compliance, respondent characteristics associated with compliance, and the validation of compliance (using total acceleration data) in a field setting *(RQ1* to *RQ4)*. The dataset of the lab study *(data source 2)* serves for evaluating squat performance; i.e., the number of performed squats counted by the experimenter *(RQ5)*.

5.2.1.1 Data Source 1: Cross-Sectional Study

This cross-sectional study was conducted by the survey company Respondi in Germany in September and October 2018. Respondi drew a quota sample from their opt-in panel based on age, education, and gender, resulting in a 3×3×2 quota plan. The company invited respondents by email. The email included an invitation to take part in the survey, an instruction to use a smartphone for survey completion, and a URL link that directed respondents to the smartphone survey. Once there, an introductory page informed respondents about the procedure of the survey and that their data would be treated confidentially. Our study was part of a larger survey with several unrelated studies and was located in the last quarter of the survey.

A total of 1,172 respondents participated in the survey. Some respondents were ineligible because they only visited the title page or they broke-off the survey before being asked any study-relevant questions ($n = 197$). In total, $n = 975$ respondents remained for statistical analyses. Another 27 respondents were excluded because there were some

technical difficulties with the acquisition of the total acceleration data. Therefore, n = 948 respondents remained for the validation of their squat task compliance by using total acceleration data.

These respondents were aged 18 to 70 years old, with a mean age of 48.0 (SD = 15.2), and 42.5% of them were female. In terms of education, 43.6% had graduated from a lower secondary school (low education level), 25.2% from an intermediate secondary school (middle education level), and 31.2% from a college preparatory secondary school or university (high education level).

5.2.1.2 Data Source 2: Lab Study

In February 2020, we conducted an additional lab study in Utrecht to get reference data on respondents' squat performance using total acceleration. At this lab study, an experimenter observed and validated respondents' task (or squat) performance. Similar to the cross-sectional study *(data source 1)*, respondents were asked to perform squats for one minute, while collecting the total acceleration of their smartphones. The experimenter observed respondents' compliance with the squat task and manually counted the number of squats that respondents performed.

Data were obtained from ten adult respondents aged 26 to 63 years, with a mean age of 33.4 (SD = 11.4), and 50.0% of them were female. In terms of education, all respondents graduated from a college preparatory secondary school or university (high education level). All respondents volunteered willingly and were familiar with the overseeing experimenter.

5.2.2 Survey Questions and Task Instructions

5.2.2.1 Data Source 1: Cross-Sectional Study

We employed 15 questions that dealt with respondents' fitness level (5 questions), general health (1 question), and physical functioning (9 questions). These questions were adopted from the Short Form (36) Health Survey [SF-36] (Ware & Sherbourne, 1992) and from a study by Keith, Stump, and Clark (2012). We also asked about respondents' body weight (1 question) and body height (1 question) to determine their Body Mass Index (BMI). All questions were presented with vertically aligned response scales and radio buttons (see Appendix A for English translations of all questions and response categories).

After the questions on fitness level, general health, physical functioning, and body weight and height, we asked respondents' about their hypothetical compliance with a fitness task during survey participation. More specifically, we asked the following question with 'Yes, I could imagine" and 'No, I could not imagine' as response categories: "*In general, could you*

imagine participating in a fitness task during a survey?"

We then asked respondents to actually do squats for one minute while holding their smartphone at chest level. To avoid an artificially sounding instruction, we slightly adapted the request for respondents who initially indicated that they would not comply with a fitness task or who did not provide an answer at all. All respondents received the opportunity to refuse their participation in the squat task by providing a reason for non-compliance in an open answer box. Complying respondents were directed to a survey page displaying a timer counting down from 60 sec to 0 seconds. Finally, we asked respondents how many squats they did by providing an open answer box to enter the number of squats. All questions and instructions were in German, which was the mother tongue of 94.2% of the respondents. To improve survey completion and task performance, we used an optimized survey layout that avoids horizontal scrolling. Figure 5.1 displays screenshots for hypothetical compliance, observed compliance including squat instruction, and the timer page for doing squats.

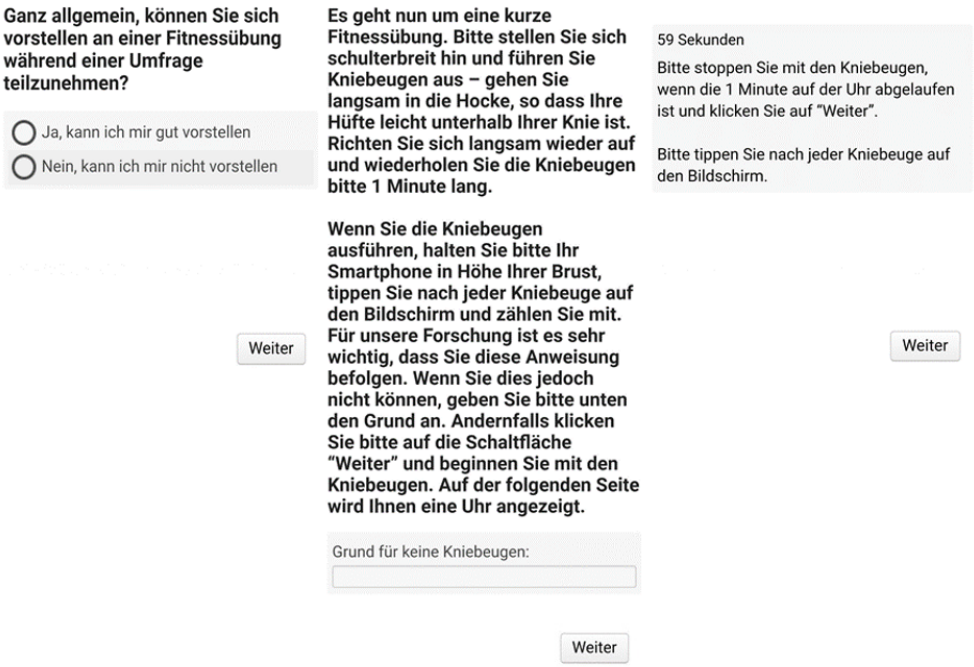


Figure 5.1 Screenshots for Hypothetical Compliance (on the Left), Observed Compliance Including Squat Instruction (in the Middle), and the Timer Page for Doing Squats (on the Right).
Note. The German versions of all questions and instructions are available from the second author on request.

5.2.2.2 Data Source 2: Lab Study

Similar to the cross-sectional study *(data source 1)*, respondents in the lab study were asked to perform squats for one minute. The design of the web survey was identical to the one of the cross-sectional study (see Figure 5.1). One important difference is that, in the lab study, respondents were asked to do the squats in four different ways, varying their intensity. This was done to ensure variation in the quality and number of squats, emulating real-world variation that is caused by respondents' motivation and skills. The four conditions were as follow:

1. Deep squats at high pace (high intensity).
2. Easy squats at high pace (medium intensity).
3. Deep squats at slow pace (medium intensity).
4. Easy squats at slow pace (low intensity).

In order to minimize the occurrence of order effects respondents conducted the four different types of squats in a randomized order. In addition, respondents were able to take breaks between each course of squats to ensure physical endurance. Due to some technical difficulties total acceleration data could not be accurately collected for four out of 40 trials, leaving us with 36 trials for statistical analyzes. The lab study contained no additional survey questions for the respondents, except for some socio-demographic questions.

5.3 Analytical Strategy

We use *data source 1* for research questions 1 to 4 and *data source 2* for research question 5.

Research Question 1. To investigate our first research question on the hypothetical and observed compliance of respondents with doing squats for one minute, we start by determining respondents' hypothetical compliance. For this purpose, we look at the proportion of respondents saying 'Yes, I could imagine' when they were asked whether they can imagine participating in a fitness task. In a next step, we determined respondents' observed compliance by looking at the proportion of respondents that did not enter any reasons in the open answer box for non-compliance when they were asked to do squats for one minute. In these cases, we assumed that respondents comply with the instructions, keeping in mind that not providing a reason does not constitute strong proof of compliance. In order to test for differences between hypothetical and observed compliance we conducted a chi-squared test.

Research Question 2. To investigate our second research question on respondents' reasons for non-compliance we coded respondents' stated reasons for non-compliance. We classified the open responses into six categories following the example of Höhne et al.

(2020).

Respondents' stated reasons for non-compliance were coded by two coders. To estimate inter-coder reliability about 13% of the reasons were coded by both coders. Then, we computed Cohen's κ to determine the agreement between the two coders. There was excellent agreement with a Cohen's $\kappa = .85$.

Research Question 3. With respect to our third research question on respondent characteristics that are associated with respondents' compliance, we conducted a logistic regression with observed compliance (1 = yes) as binary dependent variable.

To our best of knowledge there are (almost) no empirical studies investigating respondents' compliance with fitness tasks in general and squats in particular. Thus, there is little knowledge on what characteristics affect respondents' compliance. An exception is Sakshaug et al. (2010) who show that respondents' compliance with fitness tasks highly depends on health status. We therefore include the following health-related variables as independent variables: fitness level, general health, physical functioning, and BMI.

We determined respondents' fitness level using five questions asking how they assess their overall fitness level, endurance, sprint speed, strength, and flexibility. These questions were asked with completely verbalized, five-point rating scales running from 1 "Very good" to 5 "Very bad". For statistical analyses, we recoded the scales of all questions so that they run from 1 "Very bad" to 5 "Very good". An explanatory factor analysis with a principal factor method and a Promax rotation revealed that all five questions load on one factor that we call fitness level. We saved Bartlett factor scores with higher scores indicating a higher fitness level and used these scores in the logistic regression model. The fitness level factor explained 60.8% of the variance with a Cronbach's $\alpha = .88$.

In order to measure respondents' general health, we employed one self-report question that is frequently asked in health-related surveys, such as the 36-item Short Form Health Survey (SF-36) and the HRS (2018). More specifically, respondents were asked how they rate their general health with a completely verbalized, five-point rating scale running either from 1 "Excellent" to 5 "Bad", or from 1 "Bad" to 5 "Excellent" (the question was part of a scale direction experiment). For statistical analyses, we coded the scale so that it runs from 1 "Bad" to 5 "Excellent".

We determined a physical functioning score following the scoring scheme proposed by the SF-36 developers (Hays, Sherbourne, & Mazel, 1993). More specifically, scores for each of the nine questions are transformed into a scale ranging from 0 (limited a lot by health) to 100 (not limited at all by health). Subsequently, we calculated respondents' average score

across all nine questions (Hays et al., 1993). These questions were asked with completely verbalized, three-point rating scales using the following response categories: 1 “Yes, limits me greatly”, 2 “Yes, limits me somewhat”, and 3 “No, limits me not at all”.

Finally, we calculated the BMI based on respondents’ body weight (in kilogram; kg) and body height (in meters; m) that they were asked to provide. The two questions used an open answer box for entering the body weight and body height, respectively. The BMI is defined as the body weight divided by the square of the body height. Its system unit is kg/m^2.

In addition, we included several socio-demographic control variables in the logistic regression model: Female (1 = yes), age (in years), and education with high as reference: low (1 = yes) and middle (1 = yes). For the logistic regression, we calculate and report Average Marginal Effects (AMEs) and transform them to percentages to facilitate interpretation.

Research Question 4. To answer our fourth research question on the validation of respondents’ compliance using the total acceleration data, we plotted the course of total acceleration of respondents on the survey page with the timer for doing squats for one minute. In the plots, the x-axis represents the acceleration measurements over time (in milliseconds) and the y-axis represents the total acceleration measured in meter per second squared (m/s^2). In a next step, we coded the total acceleration plots and divided them into the following three categories: non-compliance, partial compliance, and full compliance. This was done for all respondents who did not provide a reason for non-compliance.

Again, the total acceleration plots were coded by two coders. To estimate inter-coder reliability about 11% of the plots were coded by both coders. Then, we computed Cohen's κ to determine the agreement between the two coders. There was excellent agreement with a Cohen's $\kappa = .86$.

In addition, we checked respondents’ time on the survey page with the timer. Four respondents who were coded as full compliers based on their plots, were subsequently coded as partial compliers because they left the survey page for doing squats before the timer was at zero.

To test for differences in average total acceleration between the categories of respondents (i.e., non-compliance, partial compliance, and full compliance) we conducted a Welch one-way test using the Games-Howell post-hoc correction procedure for unequal variances. We used the Welch one-way test and Games-Howell post-hoc procedure

because the homogeneity of variances assumption was violated (Levene's test: $F(2,460) = 104.12$, $p < .001$) and these tests do not require homogeneity of variances.

Research Question 5. To answer our final research question on the validation of squat performance (i.e., the number of squats respondents conducted), we correlate the number of squats counted by the experimenter with respondents' average total acceleration while doing squats for one minute in a lab setting *(data source 2)*. We calculated a Pearson correlation coefficient. This is done to see whether and to what extent the two measures line up. In doing so, we follow Rowlands et al. (2018) who have shown that respondents' average total acceleration correlates with their performance on chair stands (a task that is similar to ours).

5.4 Results

5.4.1 Research Question 1: Hypothetical and Observed Compliance

With respect to hypothetical compliance we found that 57.7% of the respondents could imagine taking part in a fitness task during web survey completion. Interestingly, we found that observed compliance is somewhat higher. Overall, 60.7% of the respondents stated compliance with doing squats for one minute. The result of a chi-squared test reveals that observed compliance is significantly higher than hypothetical compliance [$\chi^2(1) = 102.03$, $p < .001$].

5.4.2 Research Question 2: Reasons for Observed Non-Compliance

To answer our second research question, we investigated respondents' stated reasons for non-compliance with the squat task. As shown in Table 5.1, respondents' stated reasons for non-compliance were largely related to health issues. About 70% of the respondents who did not comply with the squat task reported health-related issues, such as having arthrosis or being injured. Another 11% of respondents reported surrounding issues, such as being in a (public) transportation vehicle or a café. The remaining 20% reported situational issues (about 4%), such as taking care of a child, reported other reasons (about 4%), such as "it is too late", reported nonsense (about 5%), such as "Vfygbvh", or refused their compliance without providing a reason (about 8%).

Table 5.1 Reasons for Non-Compliance with the Squat Task.

Reasons for non-compliance	Percentage (frequencies)
Health issues	68.7 (263)
Surrounding issues	10.7 (41)
Situational issues	3.9 (15)
Other reasons	3.7 (14)
Nonsense	5.0 (19)
Refusals	8.1 (31)

Note. Because of rounding the percentages do not add up to 100%.

5.4.3 Research Question 3: Predicting Observed Compliance

In order to investigate our third research question, which investigates respondent characteristics that are associated with squat task compliance, we conducted a logistic regression with observed compliance (1 = yes) as the dependent variable. Table 5.2 displays the results in the form of Average Marginal Effects (AMEs) and Standard Errors (SEs). Following the pseudo R^2 by Nagelkerke, the explained variance of the logistic regression model is .23.

Taking a closer look at Table 5.2 it can be observed that all health-related variables are significantly associated with observed compliance. The only exception is fitness level, which does not significantly predict observed compliance. Both general health and physical functioning show a positive association with observed compliance implying that respondents with a higher general health or physical functioning have a higher compliance propensity. The probability of complying with the squat task increases about 8.6% when general health increases one level and about 0.5% when physical functioning increases one point. In contrast, BMI shows a negative association implying that respondents with a lower BMI have a higher compliance propensity. The probability of complying with the squat task decreases about 0.9% when BMI increases by one point. Low education is the only socio-demographic variable that is significantly associated with observed compliance. The compliance probability decreases about 9.3% for low educated respondents (compared to high educated respondents).

Table 5.2 Logistic Regression of Observed Compliance with the Squat Task.

Independent variables	AME	SE
Fitness level	-0.56	2.29
General health	8.61***	2.33
Physical functioning	0.51***	0.00
BMI	-0.89**	0.33
Age	-0.11	0.12
Female	-2.80	3.57
Education with high as reference		
Low	-9.29*	4.27
Middle	-4.73	4.91
Nagelkerke's R2	.227	

Note. *p < .05, **p < .01, ***p < .001. Dependent variable: Observed compliance. See "Analytical Strategy" for the coding of all variables. AME = Average Marginal Effect. SE = Standard Error. Intercept is not significant.

5.4.4 Research Question 4: Validating Compliance with the Squat Task

To answer our fourth research question, we investigated whether we could validate respondents' compliance using the total acceleration data of the squat page. For this purpose, we coded the total acceleration plots of the survey page on which respondents were required to do the squats for one minute. We only used respondents who complied with the task by not providing a reason for non-compliance when they were asked to do so.

Based on their total acceleration plots, we assigned respondents to one out of three compliance categories: Non-compliance, partial compliance, and full compliance. Figure 5.2 displays example total acceleration plots from three respondents. These plots illustrate the total acceleration of respondents' smartphones while they were required to do squats for one minute. Total acceleration values lower than 1 indicate no motion (see Höhne & Schlosser, 2019) and, thus, non-compliance with the squat task. Following this notion, the plot on the left side indicates non-compliance, the plot in the middle indicates partial compliance, and the plot on the right side indicates full compliance.

The results of the coding of the total acceleration plots reveal that the majority of respondents partially (29.5%) or fully complied (42.2%) with the squat task when they agreed to do so. However, there is a substantial minority of respondents who did not

comply with the squat task at all (28.3%).

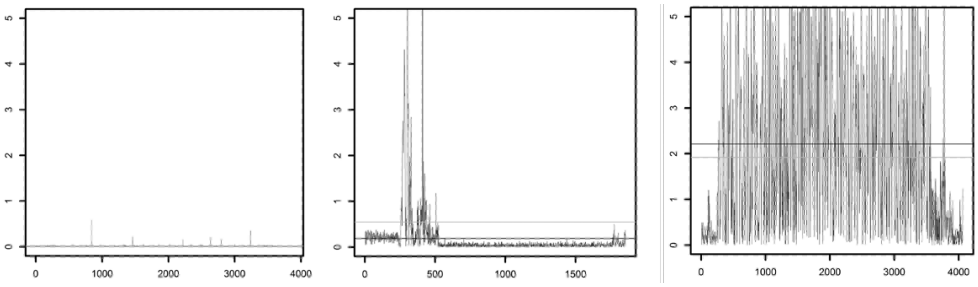


Figure 5.2 Three Example Total Acceleration Plots from Three Different Respondents.
Note. While the x-axis represents the acceleration measurements over time (in milliseconds), the y-axis represents the total acceleration measured in meter per second squared (m/s^2). The plot on the left side refers to the non-compliance category, the plot in the middle refers to the partial compliance category, and the plot on the right side refers to the full compliance category.

We also tested for total acceleration differences between the compliance categories conducting a Welch one-way test. Figure 5.3 displays the average total acceleration for the three compliance categories. The results of the Welch one-way test reveal a significant main effect across the three compliance categories [$F(2,460) = 256.62$, $p < .001$]. The results of a subsequent post-hoc comparison using the Games-Howell procedure indicate significant mean differences between the three compliance categories, except between the non-compliance and the partial compliance categories.

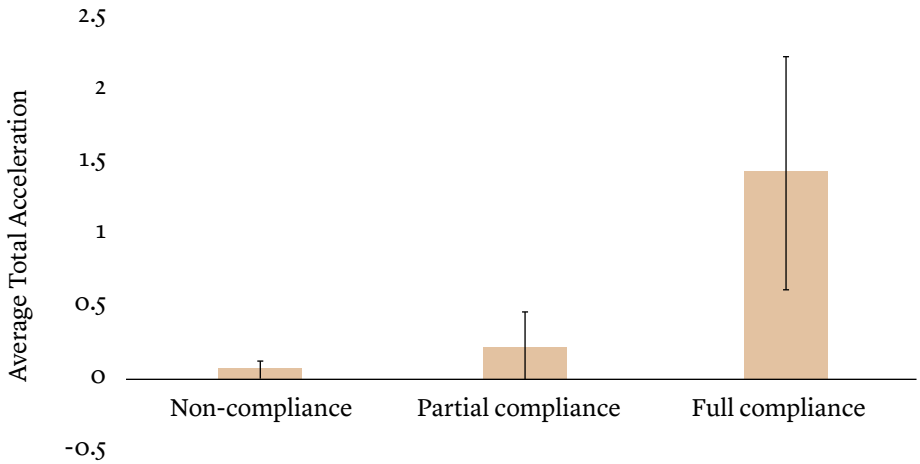


Figure 5.3 Bar Chart of the Average Total Acceleration for the Three Compliance Categories.
Note. The vertical lines within the bars represent the standard deviations.

5.4.5 Research Question 5: Validating Performance of the Squat Task

To answer our final research question on the validation of respondents' squat task performance we used data from the lab study *(data source 2)*. More specifically, we correlated the number of squats counted by the experimenter with respondents' total acceleration data. Pearson's r coefficient indicates a high and significant correlation between these two measurements [$r = 0.77, p < .001$]. This provides supporting evidence that total acceleration data can be used to validate the performance on (or the number of) squats in fitness tasks during self-administered smartphone surveys.

5.5 Discussion

The aim of this study was to investigate the feasibility of fitness tasks in self-administered smartphone surveys. More specifically, we investigated the compliance with and performance on a fitness task asking respondents to do squats for one minute while collecting high-frequency acceleration data of their smartphones. Our overall findings suggest that such fitness tasks are a feasible endeavor in self-administered smartphone surveys.

With respect to our first research question on differences between hypothetical and observed compliance, we found that observed compliance was significantly higher than hypothetical compliance. This finding differs from findings reported in other studies (Struminskaya et al., 2020a; Struminskaya et al., 2020b). In our opinion, there are two possible explanations for this phenomenon: First, usually respondents answer survey questions by selecting a response category from a predefined list. This also applies to the smartphone survey in which this study was implemented. Conducting fitness tasks during web surveys is rather seldom and might thus be an interesting and exciting task for respondents. This may lead them to participate, even though they did not intend to do so when asked hypothetically. Future studies may investigate this phenomenon further by asking respondents to outline their motivation for observed compliance. Second, the way of asking ("*In general, could you imagine participating in a fitness task during a survey?*") may have affected our results. Respondents may not have perceived this as a real question for hypothetical compliance, but more as an imagination question. In addition, the request for observed compliance was a comparatively intense and guiding question ("*If you are not able to do so, please state the reason below*"), pushing respondents towards compliance and participation. Future studies could further experiment with different ways of asking for hypothetical and observed compliance in order to optimize compliance questions and increase compliance rates.

Regarding our second research question on potential reasons for non-compliance, we

found that the majority of respondents (about 80%) gave reasons related to health, surrounding, or situation. Overall, this finding corresponds to findings reported by Höhne et al. (2020), who found that about two thirds of the respondents who did not comply with simple motion tasks reported either health-, surrounding-, or situation-related issues. Note that the comparatively high physical demands of our fitness task may have driven the high prevalence of health-related reasons for non-compliance stated by respondents. Less intensive tasks may cause fewer respondents to refuse compliance with the task because of health-related reasons.

Our third research question dealt with respondents' characteristics that are associated with compliance. In line with previous research, we found that particularly health-related variables affect compliance propensities. Respondents with a lower general health, a lower physical functioning, and a higher BMI are less likely to comply with our fitness task. These respondents may be willing but not able to comply in a squat task. As noted earlier, compliance with less physically demanding tasks than the squat task in our study may result in different correlates of compliance.

With respect to our fourth research question on validating fitness task compliance, we indeed found supporting evidence that acceleration data of smartphones can be used to validate respondents' task compliance in self-administered web surveys. Interestingly, the acceleration data showed that not all respondents who stated compliance (or did not provide a reason for non-compliance) actually complied with doing squats for one minute. Plotting the course of acceleration data over time reveals that some respondents did not comply at all or only complied partially. Nevertheless, the high observed compliance rate suggests that most respondents comply with a squat task if they agreed to do so. This indicates the general feasibility of fitness tasks in self-administered smartphone surveys to draw conclusions about respondents' physical fitness level.

Finally, regarding our fifth research question on validating the performance on fitness tasks, we found further supporting evidence that acceleration data of smartphones can be used to validate respondents' fitness task performance (i.e., number of squats) in smartphone surveys. This allows us to draw conclusions about the number of squats respondents did. Self-reports of respondents' squat performance probably suffer from an over-reporting of the number of squats due to social desirability. Additionally, using respondents' acceleration data may reduce measurement error. Fitness tasks can thus be used as a more objective supplement to health and physical fitness measures in smartphone surveys.

Our study has some limitations that provide avenues for future research. First, the fitness task was positioned close to the end of the survey. Respondents' compliance might be

higher if the task was placed earlier in the survey. Future research could vary the position of the fitness task in the survey (beginning, middle, end) in order to optimize compliance rates. Second, even though we can validate respondents' performance by counting the number of squats, we cannot make a distinction between good, deep squats, and fast, easy squats yet. Further analyses and more information on the direction of movement could help identifying the quality of the squats. Third, the samples were drawn from an access panel (cross-sectional study) and a volunteer sample (lab study). A probability sample would allow to draw more robust conclusions on fitness task compliance and performance in the general population.

In sum, this study contributes to fitness and health research by proposing a new method to study respondents' physical fitness level. So far, our results indicate that it is feasible to ask respondents to engage in fitness tasks in self-administered smartphone surveys. This increases opportunities for large surveys (e.g., HRS, SHARE, and ELSA) to switch from interviewer-administered surveys to self-administered surveys. We show that compliance with and performance on fitness tasks in self-administered smartphone surveys can be validated with acceleration data. This is much more time- and cost-efficient than employing interviewers and reduces respondent burden because respondents can complete surveys and do fitness tasks without time restrictions. We see a lot of potential for future research employing fitness tasks in self-administered smartphone surveys and extending our task (doing squats) with other commonly used tasks in public health research.

5.6 Appendix

Instruction and Question Wording Cross-Sectional Study.

Fitness level
Introduction text: The following questions are about your physical fitness level.

How would you assess your overall fitness level?
How would you assess your endurance?
How would you assess your sprint speed?
How would you assess your strength?
How would you assess your flexibility?
1 very good – 5 very bad
1 very bad – 5 very good
Recoded for analyses into 1 very bad – 5 very good

General health
In general, how would you rate your health?
1 bad – 5 excellent
1 excellent – 5 bad
Recoded for analyses into 1 bad – 5 excellent

Physical functioning
Introduction text: Does your health now limit you in the following activities?

Moderate activities, such as moving a table or pushing a vacuum cleaner.
Vigorous activities, such as running or lifting heavy objects.
Lifting or carrying groceries.
Climbing one flight of stairs.
Climbing several flights of stairs.
Bending, kneeling, or stooping.
Walking more than 100 meters.
Walking more than a kilometre.
Bathing or dressing yourself.
1 – yes, limits me greatly. 2 – yes, limits me somewhat. 3 – no, limits me not at all.

BMI
How tall are you?
Please enter your height in meters (m), e.g., 1.76 m.
Open answer box.

How much do you weigh?
Please enter your weight in kilogram (kg), e.g., 81.7 kg.
Open answer box.

Hypothetical compliance
In general, could you imagine participating in a fitness task during a survey?
1 yes, I could imagine – 2 no, I could not imagine

Observed compliance
[Hypothetical compliance "yes, I could imagine":] It is now about a short fitness task.
[Hypothetical compliance "no, I could not imagine" or missing:] Irrespective of your prior response, we would like to ask you kindly to participate in the following exercise.

Please stand shoulder wide and perform squats – crouch slowly – so that your hip is slightly below your knees. Straight slowly up and repeat the squats for 1 minute.

When doing the squats, hold your phone at chest level, tap the screen after each squat and count the squats you do. For our research it is very important that you follow these instructions. However, if you are not able to do so, please state the reason below. Otherwise, please click on the "Next" button and start with the squats. On the following page, you will see a timer that counts down.

Reason for not being able to do the squats: *[Open answer box:]*

Timer page
[Timer counting down from 60 to 0 seconds]

Please stop with the squats when the 1 minute on the timer has expired and click "Next".

Please tap on the screen after each squat.

Number of Squats
How many squats did you do?

Please enter the number of squats:
Open answer box.

Note. The original German wordings of all questions are available from the second author on request.

Chapter 6

Summary and Discussion

6.1 Summary

Smartphones have a large potential for improving data collection by using research apps and collecting sensor data. This brings opportunities to enhance or extend measurement and to simplify the response task for respondents. Sensor data can (partly) replace survey questions, and these sensors potentially generate better data than respondents can provide themselves. This seems very promising, but many methodological questions arise related to representation and measurement in smartphone surveys; are respondent willing and able to participate and share sensor data, and how useful are the additional data collected via sensors and apps?

In this dissertation we investigated the effect of smartphone surveys in terms of reducing (or enlarging) TSE components. We looked at the following TSE components in the different chapters; consent in Chapter 2, nonresponse (bias) in Chapter 3, measurement error in Chapter 4, and both measurement and representation error in Chapter 5.

In Chapter 2 we investigate how to optimize the consent question to data linkage. There is a large variability in consent rates among previous studies that cannot be fully explained yet. We investigate modifiable aspects of the consent question that increase or decrease consent rates. Our results show that how researchers ask consent questions matters greatly for consent rates. Providing respondents with arguments for data sharing always increases consent rates compared to not giving an argument. Sponsorship by a university or non-profit organization, higher incentives, using interviewers, position at the beginning of the questionnaire or in the context of a particular survey question, higher study relevance, shorter study duration and the possibility to (later) opt-out all lead to higher consent rates. We also find that there is a relatively small body of evidence for how we should ask study participants for consent to data linkage and that more research on ways to frame the consent question is needed. But we also show that there are definitely things researchers can do to increase consent probabilities, so we advise researchers to put effort in the consent request.

In Chapter 3 we investigate nonresponse and nonresponse bias at different stages of a smartphone Time Use Survey. The smartphone app contains many different stages: 1) accept an invitation to participate and install an app, 2) fill out a questionnaire on the web, 3) participate in the smartphone time use diary, 4) answer pop-up questions, and 5) share sensor (i.e. GPS and sensor) data. At every stage, some respondents fail to participate. In this study we look at all stages to get insight in the complete smartphone survey process. Our results show that cumulative nonresponse is very large, but comparable to traditional offline time use surveys. In addition, every stage introduces different selectivity. Younger respondents, who are familiar with smartphones are more likely to participate in the

smartphone parts. Lastly, respondents who participate in all stages are different from respondents that do not complete all stages: respondents who participate in all stages work more, and spend less time watching TV than the (partial) non-respondents. This means that there appears to be nonresponse bias on the variable of interest, which is problematic for making inferences. To conclude, there is a long way to go before we can use smartphones as the sole data collection mode in general population studies. This study proves that smartphone surveys are promising tools for social research, but that there is still a lot of work to do on increasing smartphone familiarity and on convincing people to do survey related tasks on smartphones.

In Chapter 4, we investigate whether the passive collection of location tracking data is sufficient to automatically establish functional locations in time use research. Functional locations can be a valuable addition to time use research, as we get an insight not just into what people are doing but also where they are doing it. This enriches measurement by allowing time use data to be understood in its context. We propose a method for analyzing GPS data and integration the location and survey data. Despite our efforts to align the time use diaries and geographical locations over time there are large differences in the time people appear to be at home, in transit or at school/work. We can only to some extent derive functional locations automatically (i.e. home location), at least for Android users. For iOS users, the results of the automated coding of functional locations are rather disappointing. Measurement error both in location tracking, also called positioning errors, and the time use diaries proves to be a major issue that makes it really hard to align both data sources. Respondents' self-reported information is thus still necessary to establish functional locations. Future research could explore different methods of filtering positioning errors and recording GPS coordinates, and variations in recording frequency length, to investigate its effect on data quality. Location tracking data can really add value to smartphone survey data collection, but it is still complicated to model the errors in the two sources.

In Chapter 5 we investigate the feasibility of using additional fitness tasks (i.e. doing squats) in a self-administered smartphone surveys. When it comes to measuring physical fitness level, most researchers rely on self-report questions employed in surveys which are prone to systematic measurement errors. Replacing self-report questions with more objective measures on respondents' physical fitness level may decrease these systematic measurement errors. Therefore, we propose a new method to study respondents' physical fitness level. Our results indicate that it is feasible to ask respondents to engage in fitness tasks in self-administered smartphone surveys. Surprisingly, observed compliance is higher than hypothetical compliance, which may be caused by the way of asking for both compliance types. Most non-compliers give health-related reasons and health-related variables affect compliance propensities: respondents may be willing but not able to do

squats. Respondents with a lower general health, a lower physical functioning, and a higher BMI are less likely to comply with the fitness task. Respondents may be willing, but not able to comply in the squat task: using less physically demanding tasks in future studies may result in (even) higher compliance rates and a more complete picture of respondents' physical fitness level. Finally, we show that we can validate respondents' compliance and performance using smartphone's total acceleration data. Using respondents' acceleration data may reduce measurement error compared to self-reports that suffer from social desirability or inaccurate recall. Fitness tasks can thus be used as a more objective supplement to traditional physical fitness measures in surveys.

6.2 Discussion

In this section we will discuss the effect of smartphone surveys in terms of reducing or enlarging TSE components and discuss avenues for future research. Just like in the introduction, we'll make a distinction between representation and measurement here.

6.2.1 Representation

How do smartphone surveys affect the TSE on the representation of the population? It seems that many respondents are willing to participate in smartphone studies, to share sensor data and to participate in additional tasks. However, our results also show that nonresponse bias occurs. Smartphone surveys involve more steps on the part of the respondent before actual participation is possible (e.g. download an app, share additional data, answer pop-up questions). Respondents need to make more effort in smartphone surveys than in traditional online surveys and at every step there is a risk that respondents drop out. There is selectivity in who participates and who does not, which introduces some bias in the representation.

There are some positive factors that may reduce this type of representation bias in the future though. Smartphone penetration and familiarity are increasing. As shown in Chapter 3, but also in many other studies, smartphone familiarity is an important predictor for (willingness to) response in smartphone surveys. As this trend is likely to continue, it may increase respondents' willingness to participate. Another positive factor is that, as shown in Chapter 2, we can quite easily improve the way we ask for consent and thereby (hopefully) increase consent rates and decrease consent bias in the future. More research on this topic is needed, but it seems that if researchers put sufficient effort and thoughts in the design of the consent request, we can probably prevent a part of the non-consent.

We would like to highlight some other, positive perspectives on the development of representation error in smartphone surveys. These perspectives all signal a potential

decrease in TSE for smartphone surveys, relative to traditional online surveys. First, smartphones offer new ways of contacting respondents. A main drawback of online surveys has always been the lack of sampling frame of email address for the general population. Smartphones may be used to overcome this problem, because they are equipped with a SIM card and can thus be contacted via a mobile phone number. This allows researchers to use random digit dialing to sample mobile phone numbers, and call or text possible respondents (Beuthner, Sand, & Silber, 2019; Couper et al., 2017). Second, we can improve coverage by using location-based surveys. As soon as a person (who has turned on its smartphone location) enters a certain location (s)he is sent a push notification to fill out a questionnaire. Generally about their recent experience regarding the site location, but the opportunities are endless. This is a very efficient way to target specific groups of people (Cardone et al., 2015; Wray et al., 2019).

For future research, we strongly advise to investigate how to make smartphone work in the real world. Most studies published so far rely on volunteers and use small samples. There are a few examples that have shown that it is possible to achieve reasonably good response rates when mobile phone app studies are conducted in the general population, but we need to test and experiment more. Second, we recommend to experiment with ways to convince people to participate. One possibility would be to use interviewers to help specific populations (e.g. old people; respondents who are not familiar with smartphones) with installing and using the research app. This would lower the threshold to participate for these groups.

6.2.2 Measurement

Whereas there are still challenges on the representation side of the TSE framework, the big promise of smartphones is in reducing measurement error. How do smartphone surveys affect the TSE on the measurement of survey concepts? Our results show that we get more objective measures and decrease measurement error when using smartphone surveys and sensor data. We show that sensor data can be used to objectively measure fitness levels, and to add context to time use research. However, sensor data are not the holy grail and measurements are not as error free as is often thought.

On the other hand, respondents' self-reports in surveys are far from perfect too. Respondents may be unable or unwilling to give certain answers. Especially in diary surveys, which respondents have to fill out for several days, measurement errors rapidly increase over time. Using sensor data to supplement survey data simplifies the response task for the respondents. For example in research diary apps, where parts of the diary can be filled out using sensor data, while respondents can influence, change and correct the data when needed. This decreases respondent burden and allows asking questions that

cannot be answered with solely sensor measures (e.g. opinions or feelings). Combining respondent and sensor data potentially increases measurement accuracy over time. Both survey and sensor measure different things, which makes it very valuable but also challenging to combine the two data sources.

Smartphone and sensor data collection do not directly make life easier for survey researchers. And they do not need to: problems in smartphone surveys can be (partly) overcome by cooperating with computer scientists and app builders. The skills to prepare mobile and sensor data for analysis and then analyzing the data pose more technical and data processing challenges to researchers than traditional data. There is for example a large variability in sampling frequency (how often the smartphone collects data) and sensor quality between devices, which may cause problems for the comparability between devices. Another challenge is that mobile phone software, and the way the two main Operating Systems interact with apps, changes continuously. App data collection is dependent on what Apple and Google allow for apps. Over time, the Operating Systems are becoming more restrictive in what they allow apps to do and what not to do.

For future research, we strongly advise to keep experimenting. There is a whole world of possibilities to explore; Experiments will eventually lead to a better understanding and optimal application of smartphone surveys for the general population. Lab studies or validation studies will provide more insight in what these sensors actually measure and how we can draw conclusions about substantive constructs. General population studies will provide insights in how smartphone surveys work on a large scale, and how factors like respondent behavior and phone type influence measurement quality. Besides, we need more knowledge on technical issues like data storage and aggregation, data linkage and building apps, especially for large scale projects. It may become more difficult to store all information since data sets are becoming more complex, large and privacy-sensitive. And off course we need to rethink ethics. We can collect large amounts of data, but the question is how we are going to store and use the data while ensuring privacy for our respondents. Respondents are the owners of their own personal data and should always have control over and limit the sharing of their own data. In addition, researchers should guarantee complete anonymity and think about what (results or open access data) they make publicly available to ensure that no is able to trace the data back to a specific individual.

6.3 Future Directions

What will survey research look like in ten years? Since 2010, there has been a large increase in technical possibilities of smartphone data collection, and the future will undoubtedly have many more possibilities for smartphone data collection. Besides, more and more people will have devices that can be used to collect information, either actively or

passively. Despite new opportunities, old methods will probably rarely become completely extinct. In the social sciences, researchers for example are often interested in attitudes and opinions that require the use of survey questions because these are difficult to measure with solely sensor data. For some research questions or target populations, face-to-face surveys may therefore still be optimal.

Looking ahead, we have many (more or less likely) ideas about what smartphone survey research will look like in 10 years. These ideas contain both technological developments and possibilities that follow from those developments. In terms of technological development, there will probably be even more data available on everyone than we can imagine. Devices will be better, sensors for passive sensor data will be better and connections between devices will we better. In ten year, we'll probably be using 6G or 7G all over the world.

But, what will this bring us? Maybe health research will largely profit from all new opportunities by connecting surveys, wearables, and research apps; Wearables and smartphone sensors will be used not only to track respondents' physical activities, but also to track other health indicators like heart rate and blood pressure.

Maybe the entire Dutch population has installed a Statistic Netherlands Survey app on their smartphone which can be used to randomly sample respondents for a short survey or an additional task.

Maybe new privacy laws further protect respondents and restrict the transfer of data from additional data sources; data will be saved on respondents' own device, and researchers can use algorithms to retrieve aggregated data or summary statistics from the respondent.

Maybe there will be many open-source code or do-it-yourself apps available so all researchers can do high quality smartphone surveys.

We don't know for sure what is going to happen, but we will keep dreaming and working to get there. Hopefully the future will profit from the endless opportunities of smartphone surveys, while keeping other (old) options to choose from.

References

References

References marked with an asterisk (*) indicate studies included in the systematic review.

Abraham, K., Maitland, A., & Bianchi, S. (2006). Nonresponse in the American time use survey: Who is missing from the data and how much does it matter? *Public Opinion Quarterly,* 70, 676–703.

*Al Baghal, T. (2016). Obtaining data linkage consent for children: Factors influencing outcomes and potential biases. *International Journal of Social Research Methodology, 19,* 623-643.

*Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2019). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review,* 0894439319828011.

Althoff, T., Sosič, R., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature,* 547, 336.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology, 8,* 89-119.

Anhøj, J., & Møldrup, C. (2004). Feasibility of collecting diary data from asthma patients through mobile phones and SMS (short message service): Response rate analysis and focus group evaluation from a pilot study. *Journal of Medical Internet Research, 6,* e42.

*Antommaria, A. H. M., Brothers, K. B., Myers, J. A., Feygin, Y. B., Aufox, S. A., Brilliant, M. H., ... & Jarvik, G. P. (2018). Parents' attitudes toward consent and data sharing in biobanks: A multisite experimental survey. *AJOB Empirical Bioethics, 9,* 128-142.

Antoun, C., & Cernat, A. (2020). Factors affecting completion times: A comparative analysis of smartphone and PC web surveys. *Social Science Computer Review, 38,* 477-489.

Antoun, C., Katz, J., Argueta, J., & Wang, L. (2018). Design heuristics for effective smartphone questionnaires. *Social Science Computer Review, 36,* 557-574.

Arentze, T., Dijst, M., Dugundji, E., Chang-Hyeon, J., Kapoen, L., Krygsman, S., ...

Timmermans, H., (2001). New activity diary format: Design and limited empirical evidence. *Transportation Research Record, 1768,* 79–88.

Avendano, M., Scherpenzeel, A. C., & Mackenbach, J. P. (2011). Can biomarkers be collected in an Internet survey? A pilot study in the LISS panel. *Social and behavioral research and the Internet: Advances in applied methods and research strategies,* 371-412.

*Balestra, M., Shaer, O., Okerlund, J., Westendorf, L., Ball, M., & Nov, O. (2016). Social annotation valence: The impact on online informed consent beliefs and behavior. *Journal of Medical Internet Research, 18,* 256–273.

Bauman, A., Bull, F., Chey, T., Craig, C. L., Ainsworth, B. E., Sallis, J. F., ... The IPS Group. (2009). The International Prevalence Study on Physical Activity: results from 20 countries. *International Journal of Behavioral Nutrition and Physical Activity, 6,* 21.

Belli, R. F., Shay, W., & Stafford, F. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public Opinion Quarterly, 65,* 45–74.

*Becker, M., Matt, C., & Hess, T. (2020). It's Not Just About the Product: How Persuasive Communication Affects the Disclosure of Personal Health Information. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 51,* 37-50.

*Berry, J. G., Ryan, P., Duszynski, K. M., Braunack-Mayer, A. J., Carlson, J., Xafis, V., & Gold, M. S. (2013). Parent perspectives on consent for the linkage of data to evaluate vaccine safety: A randomised trial of opt-in and opt-out consent. *Clinical Trials, 10,* 483–494.

Beuthner, C., Daikeler, J., & Silber, H. (2019). Mixed-Device and Mobile Web Surveys. Retrieved from: https://www.gesis.org/fileadmin/upload/SDMwiki/2019_mixed_beuthner_1.pdf

Beuthner, C., Sand, M., & Silber, H. (2019). *Text message invitations as a new way to conduct population wide online surveys? Biases and coverage Issues.* Presented at the General Online Research Conference, Berlin.

*Beuthner, C., Weiß, B., Silber, H., & Keusch, F. *(in press).* Consenting to Data Linkage – The Roll of the Data Domain, Framing, Device, Incentives and Respondent Characteristics.

*Bhatia, J., & Breaux, T. D. (2018). Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction, 25*. https://doi.org/10.1145/3267808

Biemer, P. P. (2011). Latent class analysis of survey error (Vol. 571). John Wiley & Sons.

Biemer, P. P. (2016). Total Survey Error Paradigm: Theory and Practice. In eds. C. Wolf, D. Joye, T.Smith, & Y.C. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 122 – 141). SAGE Publishers.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality (Vol. 335)*. John Wiley & Sons.

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,* 285-297.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., ... & Zuber, S. (2013). Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology, 42,* 992-1001.

Bouchard, C., Tremblay, A., Leblanc, C., Lortie, G., Savard, R., & Thériault, G. (1983). A method to assess energy expenditure in children and adults. *American Journal of Clinical Nutrition, 37,* 461–467.

*Boyd, A., Tilling, K., Cornish, R., Davies, A., Humphries, K., & MacLeod, J. (2015). Professionally designed information materials and telephone reminders improved consent response rates: Evidence from an RCT nested within a cohort study. *Journal of Clinical Epidemiology, 68,* 877–887.

*Brelsford, K. M., Ruiz, E., Hammack, C. M., & Beskow, L. M. (2019). Improving Translation and Cultural Appropriateness of Spanish Language Consent Materials for Biobanks. *Ethics & Human Research, 41,* 16-27.

*Briscoe, F., Ajunwa, I., Gaddis, A., & McCormick, J. (2020). Evolving public views on the value of one's DNA and expectations for genomic database governance: Results from a national survey. *PloS one, 15,* e0229044.

*Bryant, H., Robson, P. J., Ullman, R., Friedenreich, C., & Dawe, U. (2006). Population-based cohort development in Alberta, Canada: A feasibility study. *Chronic*

Diseases in Canada, 27, 51–59.

Buck, N., & McFall, S. (2011). Understanding Society: design overview. *Longitudinal and Life Course Studies, 3,* 5-17.

*Burstein, M. D., Robinson, J. O., Hilsenbeck, S. G., McGuire, A. L., & Lau, C. C. (2014). Pediatric data sharing in genomic research: attitudes and preferences of parents. *Pediatrics, 133,* 690–697.

Buskirk, T. D., & Andrus, C. (2012). Smart surveys for smart phones: Exploring various approaches for conducting online mobile surveys via smartphones. *Survey Practice, 5,* 1-11.

Calderwood, L., & Gilbert, E. (2018). *Measuring Young People's Physical Activity Using Accelerometers in the UK Millennium Cohort Study.* Paper presented at BigSurv18 Conference, Barcelona, Spain.

Callegaro, M. & Disogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly, 72,* 1008–1032.

Cardone, G., Cirri, A., Corradi, A., Foschini, L., Ianniello, R., & Montanari, R. (2014). Crowdsensing in urban areas for city-scale mass gathering management: Geofencing and activity recognition. *IEEE Sensors Journal, 14,* 4185-4195.

Čehovin, G., Bosnjak, M., & Lozar Manfreda, K. (2018). Meta-Analyses in Survey Methodology: A Systematic Review. *Public Opinion Quarterly, 82,* 641-660.

Chatzitheochari, S., Fisher, K., Gilbert, E., Calderwood, L., Huskinson, T., Cleary, A., & Gershuny, J. (2017). Using new technologies for time diary data collection: Instrument design and data quality findings from a mixed-mode pilot survey. *Social Indicators Research,* 1–12.

Chen, G. (2011). Mobile research: Benefits, applications, and outlooks. In P. L. P. Rau (Ed.), *Internationaliza-tion, design and global development* (pp. 11–16). Berlin, Germany: Heidelberg.

Christensen, A. I., Ekholm, O., Glümer, C., & Juel, K. (2013). Effect of survey mode on response patterns: comparison of face-to-face and self-administered modes in health surveys. *The European Journal of Public Health, 24,* 327-332.

Clement, J. (2020, November 23). *Mobile internet traffic share in selected countries 2020.* Retrieved from: https://www.statista.com/statistics/430830/share-of-mobile-internet-traffic-countries/

*Critchley, C., Nicol, D., & Otlowski, M. (2015). The impact of commercialisation and genetic data sharing arrangements on public trust and the intention to participate in biobank research. *Public Health Genomics, 18,* 160–172.

Cialdini, R. (1993). *Influence: Science and practice* (3rd ed.). New York: Harper Collins Publishers Inc.

Cloïn, M., van den Broek, A., van den Dool, R., de Haan, J., de Hart, J., van Houwelingen, P., .. . Spit, J. (2013). *Met het oog op de tijd: Een blik op de tijdsbesteding van Nederlanders* [In the interest of time: A look at the Dutch time use]. The Hague, the Netherlands: The Netherlands Institute for Social Research (SCP).

Costa, P. & McCrae, R. (1992). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological Assessment, 4,* 5–13.

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M., & Zegras, C. P. (2013). Future mobility survey. *Transportation Research Record: Journal of the Transportation Research Board, 2354,* 59–67.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In P.P. Biemer, E. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 133–154). Hoboken, NJ: Wiley

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review, 35,* 357–377.

Da Silva, M. E. M., Coeli, C. M., Ventura, M., Palacios, M., Magnanini, M. M. F., Camargo, T. M. C. R., & Camargo, K. R. (2012). Informed consent for record linkage: a systematic review. *Journal of Medical Ethics, 38,* 639-642.

*Das, M., & Couper, M. P. (2014). Optimizing opt-out consent for record linkage. *Journal of Official Statistics, 30,* 479-497.

De Bruijne, M. & Wijnant, A. (2014). Mobile response in web panels. *Social Science Computer Review, 32,* 728–742.

De Bruijne, M., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review, 31,* 482-504.

De Leeuw, E., Hox, J., & Luiten, A. (2018). International nonresponse trends across countries and years: an analysis of 36 years of Labour Force Survey data. *Survey Methods: Insights from the Field,* 1-11.

De Leeuw, E.D., Hox, J., Silber, H., Struminskaya, B., & Vis, C. (2019). Development of an international survey attitude scale: measurement equivalence, reliability, and predictive validity. *Measurement Instruments for the Social Sciences, 1,* 1-10.

De Leeuw, E. D., & Toepoel, V. (2018). Mixed-Mode and Mixed-Device Surveys. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 51-61). Cham: Springer International Publishing.

Dillman, D. (2000). *Internet and mail surveys: The tailored design method.* John Wiley, New York.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., ... & Ktori, M. (2011). Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. *PloS one, 6,* e24974.

Edwards, B. & Biddle, N. (2020). Consent to data linkage: Experimental evidence on the impact of data linkage requests and understanding and risk perceptions. *Methodology of Longitudinal Surveys 2 Monograph.* John Wiley & Sons

*Edwards, B. & Biddle, N. (in press). Consent to Data Linkage: Experimental Evidence from an Online Panel. In P. Lynn (eds.), *Advances in Longitudinal Survey Methodology.* John Wiley & Sons.

*Eisnecker, P. S., & Kroh, M. (2017). The informed consent to record linkage in panel studies: Optimal starting wave, consent refusals, and subsequent panel attrition. *Public Opinion Quarterly, 81,* 131–143

Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, B.M.S., & Toepoel, V. (2019). Where you at? Using GPS locations in an electronic Time Use Diary Study to Derive Functional Locations. *Social Science Computer Review.*

Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process? *Survey Research Methods, 13,* 195–213.

Elhoushi, M., Georgy, J., Noureldin, A., & Korenberg, M. J. (2017). A survey on approaches of motion mode recognition using sensors. *IEEE Transactions on Intelligent Transportation Systems, 18,* 1662–1686.

Ermes, M., Parkka, J., Mantyjarvi, J., & Korhonen, I. (2008). Detection of daily activities and sports with wear-able sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine, 12,* 20–26.

ESS ERIC (2018). *CRONOS Fieldwork documents.* London: ESS ERIC headquarters.

Eurostat. (2009). *Harmonised European time use surveys (HETUS): Guidelines 2008.* Luxembourg: Office for Official Publications of the European Communities.

Fan, W. & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior, 26,* 132–139

Fernihough, A. (2014). mfx: Marginal Effects, Odds Ratios and Incidence Rate Ratios for GLMs. R package version 1.1. https://CRAN.R-project.org/package=mfx

Fisher, K. & Gershuny, J. (2013). The 2014–2015 United Kingdom Time Use Survey. 10, 96–97.

*Fobia, A. C., Holzberg, J., Eggleston, C., Childs, J. H., Marlar, J., & Morales, G. (2019). Attitudes Towards Data Linkage for Evidence-Based Policymaking. *Public Opinion Quarterly, 83,* 264-279.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics, 22,* 313–328.

Gatny, H.H., Couper, M.P., & Axinn, W.G. (2013). New strategies for biosample collection in population-based social research. *Social Science Research, 42,* 1402-1409.

Gershuny, J. (2012). Too many zeros: A method for estimating long-term time-use from short diaries. *Annals of Economics and Statistics, 105/106,* 247–270.

Goldberg, L., Johnson, J., Eber, H., Hogan, R., Ashton, M., Cloninger, R., & Gough,

H. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84–96.

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science, 1,* 58-70.

*Grande, D., Asch, D. A., Wan, F., Bradbury, A. R., Jagsi, R., & Mitra, N. (2015). Are patients with cancer less willing to share their health information? privacy, sensitivity, and social purpose. *Journal of Oncology Practice, 11,* 378–383.

*Graves, A., McLaughlin, D., Leung, J., & Powers, J. (2019). Consent to data linkage in a large online epidemiological survey of 18–23 year old Australian women in 2012–13. *BMC Medical Research Methodology, 19,* 235.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly. 70,* 646–675.

Groves, R., Cialdini, R. B., & Couper, M. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56,* 475–495.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey Methodology.* John Wiley & Sons.

Groves, R. & Heeringa, S. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169,* 439–457.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74,* 849-879.

Groves, R., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly, 64,* 299–308.

Gummer, T., Quoß, F., and Roßmann, J. (2019), "Does increasing mobile device coverage reduce heterogeneity in completing web surveys on smartphones?" *Social Science Computer Review, 37,* 371-384.

Haan, M., Lugtig, P., & Toepoel, V. (2019). Can we predict device use? An investigation into mobile device use in surveys. *International Journal of Social Research*

Methodology, 22, 517–531

*Halevi, T., Kuppusamy, T. K., Caiazzo, M., & Memon, N. (2015). Investigating users' readiness to trade-off biometric fingerprint data. *2015 IEEE International Conference on Identity, Security and Behavior Analysis, ISBA 2015.*

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. *Perspectives on Psychological Science, 11,* 838-854.

Hassani, M., Kivimaki, M., Elbaz, A., Shipley, M., Singh-Manoux, A., & Sabia, S. (2014). Non-consent to a wrist-worn accelerometer in older adults: the role of socio-demographic, behavioural and health factors. *PLoS One, 9,* e110816.

Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The rand 36- item health survey 1.0. *Health economics, 2,* 217-227.

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change.* New York, NY: Routledge.

Hofman, L.F., Human Saliva as a Diagnostic Specimen, *The Journal of Nutrition, 131,* 1621S–1625S.

Höhne, J. K., Revilla, M., & Schlosser, S. (2020). Motion instructions in surveys: Compliance, acceleration, and response quality. *International Journal of Market Research, 62,* 43–57.

Höhne, J. K., & Schlosser, S. (2019). SurveyMotion: What can we learn from sensor data about respondents' completion and response behavior in mobile web surveys? *International Journal of Social Research Methodology.*

HRS Staff (2006). *HRS 2006 Final Release Codebook.* Ann Arbor, Michgian: Institute for Social Research, University of Michigan. Retrieved from: http://hrsonline.isr.umich.edu/modules/meta/2006/core/codebook/h06_00.html?_ga=2.164569977.734408553.1556538025-1847621657.1550151613

IBM Corp. (2016). *IBM SPSS statistics for Windows, version 24.0.* Armonk, NY: IBM Corp.

*Jäckle, A., Beninger, K., Burton, J., & Couper, M.P. *(in press)*. Understanding Data Linkage

Consent in Longitudinal Surveys. In P. Lynn (eds.), *Advances in Longitudinal Survey Methodology.* John Wiley & Sons.

Jäckle, A., Burton, J., Couper, M.P., & Lessof, C. (2017). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: response rates and response biases. *Institute for Social and Economic Research, University of Essex: Understanding Society Working Paper Series No. 2017-09.*

Jaspers, E., & van Tubergen, F. (2016). Children of Immigrants Longitudinal Survey in the Netherlands (CILSNL). Wave 5. Reduced version v4.0.0. DANS.

Kalter, F., Heath, A. F., Hewstone, M., Jonsson, J. O., Kalmijn, M., Kogan, I., & Van Tubergen, F. (2016). *Children of immigrants longitudinal survey in four European countries (CILS4EU)—Full version.* Data file for on-site use. GESIS Data Archive, Cologne, ZA5353 Data file Version 3.1.0.

Kantar Public Brussels (2019). *Special Eurobarometer 480. Europeans' attitudes towards Internet Security. Summary.* Retrieved from: https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/yearFrom/2019/yearTo/2020/search/internet%20security/surveyKy/2207

Kapteyn, A., Banks, J., Hamer, M., Smith, J. P., Steptoe, A., van Soest, A., ... & Wah, S. H. (2018). What they say and what they do: comparing physical activity across the USA, England and the Netherlands. *Journal of Epidemiology & Community Health, 72,* 471-476.

Keith, N. R., Stump, T. E., & Clark, D. O. (2012). Developing a self-reported physical fitness survey. *Medicine and Science in Sports and Exercise, 44,* 1388.

Keusch, F. (2015). Why do people participate in web surveys? Applying survey participation theory to internet survey data collection. *Management Review Quarterly, 65,* 183–216.

*Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public Opinion Quarterly, 83,* 210-235.

Khan, A. M., Lee, Y. K., Lee, S. Y., & Kim, T. S. (2010). Human activity recognition via an accelerometer-enabled-smartphone using kernel discriminant analysis. In

2010 5th international conference on future information technology (pp. 1-6). IEEE.

*Kim, K. K., Joseph, J. G., & Ohno-Machado, L. (2015). Comparison of consumers' views on electronic data sharing for healthcare and research. *Journal of the American Medical Informatics Association, 22,* 821–830.

*Kim, H., Bell, E., Kim, J., Sitapati, A., Ramsdell, J., Farcas, C., ... Ohno-Machado, L. (2017). iCONCUR: Informed consent for clinical data and bio-sample use for research. *Journal of the American Medical Informatics Association, 24,* 380–387.

Knulst, W. & Van den Broek, A. (1999). Do time-use surveys succeed in measuring "busyness"? Some observations of the Dutch case. *Loisirs & Société, 21,* 563– 572.

Koricheva, J., & Gurevitch, J. (2013). Place of meta-analysis among other methods of research synthesis. Retrieved from https://pdfs.semanticscholar.org/4b24/60412c23e3dfbe59536e1c9e389138660b4b.pdf

Krebs, D., & Höhne, J. K. (2019). Exploring scale direction effects and response behavior across PC and smartphone surveys. *Journal of Survey Statistics and Methodology.*

*Kreuter, F., Haas, G. C., Keusch, F., Bähr, S., & Trappmann, M. (2018). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review,* 0894439318816389.

*Kreuter, F., Sakshaug, J. W., & Tourangeau, R. (2016). The framing of the record linkage consent question. *International Journal of Public Opinion Research, 28,* 142-152.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213–236.

Lai, J., Vanno, L., Link, M., Pearson, J., Makowska, H., Benezra, K., & Green, M. (2010). Life360: Usability of mobile devices for time use surveys. *Survey Practice, 3.*

Lemay, M. (2010). *Understanding the mechanism of panel attrition.* Unpublished Doctoral thesis, Doctor of Philosophy, University of Maryland, College Park, MD.

Li, T., Puhan, M. A., Vedula, S. S., Singh, S., & Dickersin, K. (2011). Network meta-analysis-highly attractive but more methodological research is needed. *BMC medicine, 9,* 79.

Lindau, S. T., & McDade, T. W. (2008). Minimally invasive and innovative methods for biomeasure collection in population-based research. In *Biosocial Surveys.* National Academies Press (US).

Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., & Langer Tesfaye, C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly, 78,* 779-787.

Lipps, O. (2009). Attrition of households and individuals in panel surveys. *SOEPpapers on Multidisciplinary Panel Data Research, 164.* Retrieved from http://hdl.handle.net/10419/150711

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis.* Oxford University Press.

Lugtig, P. (2014). Panel attrition: Separating stayers, fast attriters, gradual attriteres, and lurkers. *Sociological Methods & Research, 43,* 699–723.

Lugtig, P. & Schouten, J.G. *(submitted)* Combining surveys and organic data from sensors in Designed Big Data. Three case studies. *Public Opinion Quarterly.*

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review, 31,* 725–743.

Mavletova, A., & Couper, M. P. (2013). Sensitive topics in PC web and mobile web surveys: Is there a difference? *Survey Research Methods, 7,* 191–205.

McCool, D., Schouten, J.G. & Lugtig, P. (in press). TABI: a smartphone travel app to produce official statistics on travel behavior in the Netherlands. *Journal of Official Statistics.*

*McGuire, A. L., Oliver, J. M., Slashinski, M. J., Graves, J. L., Wang, T., Kelly, P. A., ... Hilsenbeck, S. G. (2011). To share or not to share: A randomized trial of consent for data sharing in genome research. *Genetics in Medicine, 13,* 948–955.

Menai, M., Van Hees, V. T., Elbaz, A., Kivimaki, M., Singh-Manoux, A., & Sabia, S. (2017). Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. *ScientificRreports, 7,* 45772.

Michelson, W. (2005). *Time use: Expanding the explanatory power of the social sciences.* Boulder, Colorado: Paradigm Publishers.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7,* 221-237.

Minnen, J., Glorieux, I., Van Tienoven, T., Daniels, S., Weenas, D., Deyaert, J., . . . Rymenants, S. (2014). Modular online time use survey (MOTUS)— translating an existing method in the 21st century. *Electronic International Journal of Time Use Research, 11,* 73–93.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med, 6,* e1000097.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26,* 67–82.

Morgan, K., Page, N., Brown, R. et al. Sources of potential bias when combining routine data linkage and a national survey of secondary school-aged children: a record linkage study. *BMC Medical Research Methodology, 20,* 178.

*Nodora, J. N., Komenaka, I. K., Bouton, M. E., Ohno-Machado, L., Schwab, R., Kim, H.-E., ... Martinez, M. E. (2017). Biospecimen sharing among Hispanic women in a safety-net clinic: Implications for the precision medicine initiative. *Journal of the National Cancer Institute, 109.*

ONS. (2006). *The United Kingdom 2005 time use survey.* Retrieved from http://www.ons.gov.uk/ons/rel/life styles/time-use/2005-edition/index.html

* Pascale, J. (2011). Requesting consent to link survey data to administrative records: Results from a split-ballot experiment in the survey of health insurance and program participation. Technical Report, *Survey Methodology Series #2011-03* , United States Census Bureau. Retrieved from www.census.gov

*Passmore, S. R., Jamison, A. M., Hancock, G. R., Abdelwadoud, M., Mullins, C. D., Rogers, T. B., & Thomas, S. B. (2019). "I'm a Little More Trusting": Components of Trustworthiness in the Decision to Participate in Genomics Research for African Americans. *Public Health Genomics, 22,* 215-226.

Patel, V., Nowostawski, M., Thomson, G., Wilson, N., & Medlin, H. (2013). Developing a smartphone "app" for public health research: The example of measuring observed smoking in vehicles. *Journal of Epidemiology and Community Health, 67,* 446–452.

Pew Research Center. (2018a). Across 39 countries, three-quarters say they use the internet. Washington, DC. Retrieved from http://www.pewglobal.org/2018/06/19/across-39-countries-three-quarters-say-they-use-the-internet/

Pew Research Center. (2018b). Smartphone ownership on the rise in emerging economies. Washington, DC. Retrieved from http://www.pewglobal.org/2018/06/19/2-smartphone-ownership-on-the-rise-in-emerging-economies/

Pforr, K. (2016). *Incentives (Version 2.0).* (GESIS Survey Guidelines). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/gesis-sg_en_001

Pinter, R. (2015). *Willingness of online access panel members to participate in smartphone application-based research.* In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies (pp. 141-156). London: Ubiquity Press.

Plowman, L., & Stevenson, O. (2012). Using mobile phone diaries to explore children's everyday lives. *Childhood, 19,* 539–553

*Pratap, A., Allred, R., Duffy, J., Rivera, D., Lee, H. S., Renn, B. N., & Areán, P. A. (2019). Contemporary views of research participant willingness to participate and share digital data in biomedical research. *JAMA network open, 2,* e1915717-e1915717.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity, 5,* 56.

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An emerging tool for social scientists. *Sociological Methods & Research, 37,* 426-454.

Revilla, M., & Couper, M. P. (2018a). Comparing grids with vertical and horizontal item-by-item formats for PCs and Smartphones. *Social Science Computer Revew, 36,* 349–368.

Revilla, M., & Couper, M. P. (2018b). Testing different order-by click question layouts for PC and Smartphone respondents. *International Journal of Social Research Methodology.* Advance online publication.

Revilla, M., Couper, M. P., & Ochoa, C. (2019). Willingness of online panelists to perform additional tasks. *Methods, Data, Analyses, 13,* 29.

Revilla, M., & Ochoa, C. (2015). What are the links in a Web survey among response time, quality, and auto-evaluation of the efforts done?. *Social Science Computer Review, 33,* 97-114.

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels really need to allow and adapt surveys to mobile devices? *Internet Research, 26,* 1209–1227.

Richter, D., Körtner, J., & Saßenroth, D. (2014). Personality has minor effects on panel attrition. *Journal of Research in Personality, 53,* 31–35.

Roeters, A. (2017). *Een week in kaart (editie 1)* [Mapping a week (edition 1)]. The Hague, the Netherlands: Institute for Social Research (SCP).

Rowlands, A. V., Edwardson, C.L., Davies, M.J., Khunti, K., Harrington, D. & Yates, T. (2018). Beyond Cut Points. Accelerometer metrics that capture the physical activity profile. *Medicine & Science in Sports & Exercise, 50,* 1323-1332.

Rücker, G., Krahn, U., König, J., & Efthimiou, O. Schwarzer G. (2020) netmeta: Network Meta-Analysis using Frequentist Methods. R package version 1.2-1. https://cran.r-project.org/web/packages/netmeta/index.html

Sakshaug, J. W., Couper, M. P., & Ofstedal, M. B. (2010). Characteristics of physical measurement consent in a population-based survey of older adults. *Medical Care, 48,* 64 – 71.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods & Research, 41,* 535-569.

*Sakshaug, J. W., & Kreuter, F. (2014). The effect of benefit wording on consent to link survey and administrative records in a web survey. *Public Opinion Quarterly, 78,* 166-176.

*Sakshaug, J. W., Schmucker, A., Kreuter, F., Couper, M. P., & Singer, E. (2019b). The effect of framing and placement on linkage consent. *Public Opinion Quarterly, 83,* 289-308.

*Sakshaug, J. W., Stegmaier, J., Trappmann, M., & Kreuter, F. (2019a). Does Benefit Framing Improve Record Linkage Consent Rates? A Survey Experiment. In *Survey Research Methods, 13,* 289-304.

*Sakshaug, J. W., Tutz, V., & Kreuter, F. (2013). Placement, wording, and interviewers: Identifying correlates of consent to link survey and administrative data. *Survey Research Methods, 7,* 133–144.

*Sakshaug J.W., Wolter S. & Kreuter F. (2015), Obtaining Record Linkage Consent: Results from a Wording Experiment in Germany. *Survey Insights: Methods from the Field.* Retrieved from http://surveyinsights.org/?p=7288

*Sala, E., Knies, G., & Burton, J. (2014). Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology, 17,* 455-473.

Salthouse, T. (2014). Selectivity of attrition in longitudinal studies of cognitive functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 69,* 567–574.

*Sanderson, S. C., Brothers, K. B., Mercaldo, N. D., Clayton, E. W., Antommaria, A. H. M., Aufox, S. A., ... Holm, I. A. (2017). Public Attitudes toward Consent and Data Sharing in Biobank Research: A Large Multi-site Experimental Survey in the US. *American Journal of Human Genetics, 100,* 414–427. https://doi.org/10.1016/j.ajhg.2017.01.021

Scherpenzeel, A. (2009). Start of the LISS panel: Sample and recruitment of a probability-based internet panel. CentERdata, Tilburg. Retrieved from https://www.lissdata.nl/sites/default/files/bestanden/Sample%5C_and%C_Recruitment.pdf

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: How the

LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/ Bulletin de Méthodologie Sociologique, 109,* 56– 61.

Scherpenzeel, A. (2017). Mixing online panel data collection with innovative methods. In *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 27-49). Springer VS, Wiesbaden.

Scherpenzeel, A. & Das, J. (2010). "True" longitudinal and probability-based internet panels—research portal. In J. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet* (pp. 77–103). Boca Raton: Taylor & Francis.

Schlich, R., & Axhausen, K. W. (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation, 30,* 13–36.

Schlosser, S., & Höhne, J.K. (2018). ECSP – Embedded Client Side Paradata. *Zenodo.*

Schlosser, S., & Mays, A. (2018). Mobile and dirty: Does using mobile devices affect the data quality and the response process of online surveys?. *Social Science Computer Review, 36,* 212-230.

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R* (Vol. 4784). Berlin: Springer International Publishing.

*Shah, N., Coathup, V., Teare, H., Forgie, I., Giordano, G. N., Hansen, T. H., ... Kaye, J. (2018). Sharing data for future research—engaging participants' views about data governance beyond the original project: a DIRECT Study. *Genetics in Medicine.*

Singer, E. S., Gebler, N., Raghunathan, T., Van Hoewyk, J., & McGonagle, K. (1999). The effect of incentives on response rates in face-to-face, telephone, and mixed mode surveys: results of a meta-analysis. *Journal of Official Statistics, 15,* 217-230.

Smeets, L.S.M., Lugtig, P. & Schouten, J.G. *(submitted)* Automatic travel mode prediction in a national travel survey. JRSS:A https://github.com/LaurentSmeets/Master-Thesis

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

Sonck, N., & Fernee, H. (2013). *Using smartphones in survey research: A multifunctional tool.*

Implementation of a time use app; A feasibility study. The Hague: The Netherlands Institute for Social Research (SCP).

Song, H. Y., & Lee, J. S. (2015). Detecting positioning errors and estimating correct positions by moving window. *PLoS One, 10,* e0143618.

Statistics Netherlands (CBS). (2013). *Tijdsbstedingsonderzoek 2011/2012 - onderzoeksdocumentatie.* [time use survey 2011/2012 - research documentation]. Retrieved from https://catalogus.boekman.nl/pub/P13-0520C.pdf

Statline (CBS) (2019). *Internet; toegang, gebruik en faciliteiten.* Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83429NED/table?fromstatweb

Stocké, V. (2006). Attitudes toward surveys, attitude accessibility and the effect on respondents' susceptibility to nonresponse. *Quality and Quantity, 4,* 259–288.

Stoop, I. A. L. (2005). *The hunt for the last respondent: Non-response in sample surveys.* The Hague: The Netherlands Institute for Social Research—SCP.

Stoop, I. A. L. (2007). No time, too busy. Time strain and survey cooperation. In G. Loosveldt, M. Swijngedouw, & B. Chambré (Eds.), *Measuring meaningful data in social research* (pp. 301–314). Leuven: Acco.

Stopher, P., FitzGerald, C., & Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies, 16,* 350–369.

Stopher, P., & Shen, L. (2011). In-depth comparison of global positioning system and diary records. *Transportation Research Record: Journal of the Transportation Research Board, 2246,* 32-37.

*Struminskaya, B., Lugtig, P., Schouten, B., Toepoel, V., Haan, M., Dolmans, R., ... & Luiten, A. (2020a). Collecting smartphone sensor measurements in the general population: Willingness and nonparticipation bias. *Public Opinion Quarterly.*

*Struminskaya, B, Lugtig, P., Schouten, J.G., Toepoel, V., Giesen, D. & Dolmans, R. (2020b). Sharing of smartphone sensor-collected data: Willingness, participation, and non-participation bias. *Public Opinion Quarterly.*

Subar, A., Kirkpatrick, S., Mittle, B., Zimmerman, T., Thompson, F., Bingley, C., . . .

Potischman, N. (2012). The automated self-administered 24-hour dietary recall (ASA24): A resource for researchers, clinicians, and educators from the National Cancer Institute. *Journal of the Academy of Nutrition and Dietetics, 112,* 1134–1137.

Sugie, N. F. (2018). Utilizing smartphones to study disadvantaged and hard-to-reach groups. *Sociological Methods & Research, 46,* 458–491.

Thompson, F. E., Dixit-Joshi, S., Potischman, N., Dodd, K. W., Kirkpatrick, S. I., Kushi, L. H., ... Subar, A. F. (2014). Comparison of interviewer-administered and automated self-administered 24-hour dietary recalls in 3 diverse integrated health systems. *American Journal of Epidemiology, 12,* 970–978.

Toepoel, V. (2013). *Informing panel members about study results: Effects of traditional and innovative forms of feedback on participation.* Paper presented at the Workshop Longitudinal Research in Internet Panels, Mannheim.

Toepoel, V. (2016). *Doing Surveys Online.* London: Sage.

Toepoel, V., & Lugtig, P. (2015). Online surveys are mixed-device surveys. Issues associated with the use of different (mobile) devices in web surveys. *Methods, Data, Analyses, 9,* 155 – 162.

Toninelli, D., & Revilla, M. (2016). Smartphones vs PCs: Does the Device Affect the Web Survey Experience and the Measurement Error for Sensitive Topics? - A Replication of the Mavletova & Couper's 2013 Experiment. *Survey Research Methods, 10,* 153-169.

Tourangeau, R., & Ye, C. (2009). The framing of the survey request and panel attrition. *Public Opinion Quarterly, 73,* 338-348.

Trappmann, M., Beste, J., Bethmann, A., & Müller, G. (2013). The PASS panel survey after six waves. *Journal for Labour Market Research, 46,* 275-281.

Troiano, R. P., McClain, J. J., Brychta, R. J., & Chen, K. Y. (2014). Evolution of accelerometer methods for physical activity research. *Britisch Journal of Sports Medicine, 48,* 1019-1023.

Van Ingen, E., Stoop, I. A. L., & Breedveld, K. (2008). Nonresponse in the Dutch Time Use Survey: Strategies for response enhancement and bias reduction. F*ield Methods, 21,* 69–90.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1-48. URL: http://www.jstatsoft.org/v36/i03/

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30,* 473–483.

Wells, T., Bailey, J. T., & Link, M. W. (2013). Filling the void: Gaining a better understanding of tablet-based surveys. *Survey Practice, 6,* 1-9.

Wenz, A., Jäckle, A., & Couper, M. P. (2017). Willingness to use mobile technologies for data collection in a probability household panel. *Institute for Social and Economic Research, University of Essex: Understanding Society Working Paper Series,* (2017-10).

*Weydert, V., Desmet, P., & Lancelot-Miltgen, C. (2019). Convincing consumers to share personal data: double-edged effect of offering money. *Journal of Consumer Marketing.*

Wray, T. B., Pérez, A. E., Celio, M. A., Carr, D. J., Adia, A. C., & Monti, P. M. (2019). Exploring the Use of Smartphone Geofencing to Study Characteristics of Alcohol Drinking Locations in High- Risk Gay and Bisexual Men. *Alcoholism: Clinical and Experimental Research, 43,* 900-906.

Nederlandse Samenvatting

Nederlandse Samenvatting

Smartphones zijn niet meer weg te denken uit ons leven, maar kunnen we ze ook succesvol inzetten voor wetenschappelijk onderzoek? In 2019 had 92,1% van de Nederlanders een smartphone met internettoegang. Voor de leeftijdsgroep tussen 18 en 65 jaar was dit zelfs 95.2% (Statline, 2019). Het zal niemand ontgaan zijn dan een smartphone voor veel meer wordt gebruikt dan alleen bellen en overal mee naar toe gaat. Dit biedt mogelijkheden om smartphones te gebruiken voor wetenschappelijk onderzoek.

Smartphones hebben een groot potentieel voor het verbeteren van de survey datacollectie. Data van de sensoren in de telefoon kunnen (deels) de surveyvragen vervangen en mogelijk zelfs betere data generen dan respondenten zelf kunnen leveren. Dit lijkt veelbelovend maar er bestaan nog veel methodologische vragen: zijn respondenten bereid en in staat om deel te nemen en sensordata te delen en hoe nuttig zijn de extra gegevens die via sensoren en apps worden verzameld?

In dit proefschrift hebben we het effect onderzocht van smartphone-surveys op het verminderen of vergroten van verschillende error-componenten. Deze error-componenten kunnen worden gesplitst in errors in de meting zelf en errors in de representatie van de populatie.

In Hoofstuk 1 leggen we in meer detail uit wat de kansen en risico's zijn van smartphone surveys. Verder bevat het een beschrijving van de inhoud van de aankomende hoofd-stukken van dit proefschrift (in het Engels). Deze beschrijving volgt hieronder ook in het Nederlands.

In Hoofdstuk 2 onderzoeken we hoe we de effectiviteit van de consentvraag kunnen verhogen. In de consentvraag geven respondenten toestemming om aanvullende data te verzamelen, denk hierbij aan administratieve bronnen, sensoren of social media. Het behalen van hoge consentpercentages is essentieel om aanzienlijke hoeveelheden missing data te vermijden en het risico op bias te minimaliseren.

In Hoofstuk 2 onderzoeken wij door middel van een systematische review en meta-analyse welke aanpasbare aspecten van de consentvraag de percentages van toestemming verhogen of verlagen. Onze resultaten laten zien dat de manier waarop onderzoekers toestemming vragen erg belangrijk is voor het bereiken van hoge consentpercentages. De resultaten van deze studie kunnen nuttig zijn bij het ontwerpen van toekomstige onderzoeksprotocollen.

In Hoofdstuk 3 onderzoeken we in hoeverre non-response en non-response-bias

voorkomen in een smartphone studie. Dit doen we in de smartphone-only versie van het Nederlandse Tijdsbestedingsonderzoek (TBO). Dit onderzoek werd uitgevoerd met een smartphone tijdsbestedingsdagboek app en verspreid in het Nederlandse LISS-panel. De smartphone app uit deze studie bevat veel verschillende stappen (bijv. het invullen van vragenlijsten, het installeren van de app, het invullen van het dagboek, het beantwoorden van pop-up vragen, het delen van GPS-locaties) die variëren in hoe indringend en tijdsintensief ze zijn. In elke fase kunnen respondenten besluiten om niet mee te doen. Onze resultaten tonen aan dat in elke fase respondenten uitvallen en dat de geaccumuleerde non-response zeer groot is, maar vergelijkbaar met de traditionele offline TBO's. Bovendien introduceert elke fase een andere selectiviteit en heeft deze selectiviteit invloed op de tijdsbestedings-schattingen: er treedt dus non-response bias op.

In Hoofdstuk 4 richten we ons op meetfouten bij het verzamelen van sensordata in een smartphone-vragenlijst. We beoordelen of de passieve verzameling van de geografische locaties (GPS coördinaten) van de deelnemers voldoende is om hun functionele locaties (bijv. thuis, werk of onderweg) vast te stellen. We stellen een nieuwe methode voor om de GPS-gegevens te analyseren, de locatie- en onderzoeksgegevens te integreren en de variabiliteit te verklaren. Ondanks grote inspanningen om de tijdsbestedingsdagboeken en geografische locaties in de tijd op elkaar af te stemmen waren er grote verschillen in de tijd dat deelnemers thuis, onderweg of op school/werk lijken te zijn. We moesten daarom concluderen dat grote meetfouten voorkomen in zowel de geografische locatie data als de dagboeken. We gaan ook dieper in op de methodologische uitdagingen van het analyseren en integreren van GPS-locatiegegevens met zelf-gerapporteerde tijdbestedingsdata.

In Hoofdstuk 5 onderzoeken we zowel de representatie als de meting in een innovatief en experimenteel onderzoek naar het gebruik van sensordata in fitness- en gezondheidsonderzoek. We stellen een nieuwe methode voor om het fysieke fitnessniveau van respondenten te meten en onderzoeken de toepasbaarheid van aanvullende fitnesstaken (d.w.z. het doen van squats of kniebuigingen) in een smartphone-studie. Onze resultaten laten zien dat het haalbaar is om respondenten te vragen om fitnesstaken uit te voeren in een smartphone-studie. De meerderheid van de respondenten is bereid om mee te doen. De meeste non-compliers geven gezondheid gerelateerde redenen aan en gezondheid gerelateerde variabelen beïnvloeden de kans op deelname: respondenten zijn misschien wel bereid, maar niet in staat om squats te doen. Vervolgonderzoek zou daarom dieper in kunnen gaan op de mogelijkheden met minder intensieve fitnesstaken. Daarnaast laten we zien dat we deelname en prestaties kunnen valideren met behulp van de totale acceleratiedata van de smartphone.

In Hoofdstuk 6 sluiten we af met een reflectie op het werk en de uitleg die gegeven is in de hoofdstukken van dit proefschrift.

About the Author and Publications

About the Author

Anne Elevelt was born on June 5th 1993 in Eindhoven, the Netherlands. In 2014 she obtained her BSc. in Pedagogical Sciences from Utrecht University. In 2016 she completed the research master Educational Sciences: Learning in Interaction and obtained her MSc. from Utrecht University.

In September 2016, Anne started as a PhD Candidate at the department of Methodology & Statistics at Utrecht University. During her PhD, she also gave a lot of presentations, by invitation and at conferences. She has taught survey methodology courses at Utrecht University, as well as short courses on smartphone surveys and webinars on mixed device use at international programs such as WAPOR and EMOS. Between 2018 and 2019, Anne was active as a PhD Representative in the PhD council of the faculty of Social Sciences of Utrecht University. In 2019, she spent two months as a Visiting Research with the German Internet Panel Team at the University of Mannheim. In 2020 she coordinated the 'Think Tank Young Power" on Corona-related issues.

As of February 2021 Anne will continue the work on the innovation of (survey) data collection at Statistics Netherlands (CBS).

Publications

Toepoel, V., & **Elevelt, A.** (2020). Mobile and Sensor Data Collection. In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), *SAGE Research Methods Foundations.* http://dx.doi.org/10.4135/9781526421036913576

Toepoel, V., Lugtig, P., Struminskaya, B., **Elevelt, A.**, & Haan, M. (2020). Adapting Surveys to the Modern World: Comparing a Research Messenger Design to a Regular Responsive Design for Online Surveys. *Survey practice, 13,* 1 – 10. https://doi.org/10.29115/SP-2020-0010

Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, B.M.S., & Toepoel, V. (2019). Where you at? Using GPS locations in an Electronic Time Use Diary Study to Derive Functional Locations. *Social Science Computer Review.* https://doi.org/10.1177/0894439319877872

Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process? *Survey Research Methods,* 13, 195–213. https://doi.org/10.18148/srm/2019.v13i2.7385

Lensvelt-Mulders, G.J.L.M., Lugtig, P., **Elevelt, A.**, Bos, P., & Helms, A. (2016). *Aan de Grenzen van het Meetbare - De Methodologische Kwaliteit van Internationale Studies naar de Omvang van aan Prostitutie Gerelateerde Mensenhandel met Nadruk op Noordwest Europa.* The Hague: WODC, Ministerie van Veiligheid en Justitie.

Dankwoord

Dankwoord

Het is zo ver; “mijn” boek is af. Cliché maar waar; ik had het niet gekund zonder jullie. En koffie. Bedankt.

Om te beginnen met mijn supervisors. Peter en Vera, jullie namen me aan als broekie in de surveywereld. Ik heb ontzettend veel van jullie geleerd. Nog belangrijker, jullie waren fantastische begeleiders. Bedankt voor jullie vertrouwen in mij en jullie optimisme als ik zelf even niet meer geloofde dat het goed kwam. Peter, bedankt voor alle steun, gezelligheid en conferentie-biertjes. Vera, bedankt voor je oprechte interesse, betrokkenheid en de conferentie-pizza’s. Ook dank voor mijn promotor, Peter van der Heijden.

Ik wil de commissie bedanken voor het lezen en beoordelen van dit proefschrift. Edith, bedankt dat je me (virtueel) meenam om samen webinars te geven, maar ook voor alle gezellige kletspraatjes over poezen en kippen. Barry, bedankt voor je hulp bij het zoeken naar een baan. Het SCP en LISS panel, en met name Ineke Stoop, Henk Fernee en Nathalie Sonck, bedankt voor de toegang tot jullie data die ik heb gebruikt voor Hoofdstuk 3. Stijn en Wim, bedankt voor de toegang tot jullie data en het NSCR en de fijne samenwerking. Jan and Annelies, and everyone else at the German Internet Panel, thanks for the great time in Mannheim. I really enjoyed all coffee and lunch breaks, informal talks and working together with you.

Collega’s van de M&S afdeling. Bedankt voor de gezelligheid en fijne werksfeer. Dankzij jullie was het de afgelopen vier jaar extra leuk om op de universiteit te werken. Ayoub, Boukje, Diede, Erik-Jan, Flip, Jeroen, Kees, Kevin, Lientje, Marieke, Oisín bedankt voor alle gezellige koffie momenten, lunches en borrels. Thanks to all my (old) C1.21 roommates. Hidde, door ons dagelijks koffiemomentje begonnen al mijn werkdagen goed. Sanne, bedankt voor alle koffie en gezelligheid. Duco, dank voor alle biertjes en telefonische wandelingen. Karlijn, we zijn tegelijk begonnen en hebben de afgelopen 4 jaar niet alleen een kamer, maar ook al onze (PhD) frustraties, WIDM-theorieën en successen gedeeld. Marieke, wat ben ik blij met jou als surveywereld- en conferentie-buddy. Laura ik ben blij dat we na onze gezamenlijke conferenties ook nog een jaartje samen hebben mogen werken. Sjoerd en Corine, ik hoop dat we heel snel weer spontaan gaan borrelen. Survey group, thanks for all outings, discussions, thinking along, and interest in my work and life.

Mijn paranimfen. Vivian en Fayette; jullie hebben mijn hele PHD achter me gestaan, en ik ben heel blij dat jullie dit letterlijk doen tijdens mijn verdediging. Vivian, jou heb ik nodig om tegen me te schreeuwen als ik zenuwachtig word. Ik ben ongelofelijk trots op wie je bent en wat je doet om je dromen waar te maken. Fayette, ik heb mijn hele PhD samen met jou beleefd. Ik geniet van je creativiteit, warmte en onze fietstochten.

Mijn vrienden. Charlotte, Angela & Bodine, Iris, Roanne, Marloes, Iris & Dennis, Nino & Mirjam, *Surfdudes & Chicas, Singlefeestje, Outliers,* en alle anderen; ik ben zo blij dat jullie er zijn! Bedankt voor de goede gesprekken, steun, gezelligheid, eerlijkheid, afleiding, humor, etentjes, spelletjes, drankjes, en carnaval. Tot slot de *InterTransparanCie;*. Jeffrey, Lars, Koen, Mirthe en Vivian, bedankt voor jullie humor, relativering, gekkigheid en voorliefde voor onveilige sfeer. Maar ook voor alle escaperooms, bierproeverijen, wandelingen, karaoke-avonden en al het ongevraagde advies. Ik hoop dat we nog 100 jaar op *InterWeekend* gaan.

Familie van Kerkhof. Peter, Thea, David, Judith, Quinten, Lotte en Gerda, bedankt voor de buitengewoon warme ontvangst, alle gezelligheid, en jullie interesse in mij en mijn proefschrift. Ik voel me ontzettend welkom bij jullie.

Mijn familie. Jullie zijn mijn basis. Bedankt voor jullie onvoorwaardelijke steun en liefde. Ik ben ongelofelijk trots op jullie. Papa en Mama, jullie hebben mij geleerd om nooit op te geven en dat alles mogelijk is. Bedankt dat ik altijd bij jullie thuis mag komen. Bram, ik hou van onze eerlijkheid, grapjes en whatsapp gesprekken. Camille, ik vind je een topper. Tineke en Peter, dank voor jullie (ouderlijke) trots en betrokkenheid. Oma, wat ben ik blij dat u hierbij kan zijn.

Thomas. Jij bent alles. Dankjewel. Voor wat je voor mijn proefschrift hebt gedaan, maar nog veel meer voor de rest: Dat je me rust geeft, ik altijd met (en om) je kan lachen, me Excel sheets laat maken, mijn “slechte” muzieksmaak deelt, ik zoveel van je leer, je zo’n coole mountainbiker bent en omdat je altijd het leven met me wil vieren. Jij bent een feestje.