

Modernization of Data Collection Methods

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von
Georg-Christoph Haas

Hauptamtlicher Dekan der Fakultät für Sozialwissenschaften:

Prof. Dr. Michael Diehl

Erstbetreuerin:

Prof. Dr. Frauke Kreuter

Zweitbetreuer:

Prof. Dr. Florian Keusch

Erstgutachterin:

Prof. Bella Struminskaya, PhD

Zweitgutachter:

Prof. Joseph Sakshaug, PhD

Tag der Disputation:

1. Juli 2021

Contents

Contents	III
List of Figures.....	VII
List of Tables	X
1 Introduction.....	11
1.1 Quality in Establishment Surveys (QuEst)	20
1.2 Project A9: Survey mode, survey technology and technology innovations in data collection	23
1.3 IAB-SMART	24
1.4 The modernization of data collection methods	29
References	30
2 Comparing the Response Burden between Paper and Web Modes in Establishment Surveys	39
2.1 Abstract.....	39
2.2 Introduction.....	39
2.3 Background.....	41
2.3.1 Benefits of web surveys.....	42
2.3.2 Conceptualizing and measuring response burden	43
2.3.3 Possible effects of paper and web surveys on response burden	45
2.3.4 Hypotheses	46
2.4 Data.....	48
2.5 Methods	54
2.5.1 Response burden variables	54

2.5.2	Independent variables.....	55
2.5.3	Models.....	56
2.6	Results.....	59
2.7	Conclusion.....	66
	References	71
	Appendix	80
3	Comparing Single-sitting Versus Modular Text Message Surveys in Egypt ...	82
3.1	Abstract.....	82
3.2	Introduction.....	83
3.3	Using a modular design in text message survey	86
3.3.1	Unit and item response	87
3.3.2	Nonresponse bias.....	88
3.3.3	Substantive responses.....	89
3.3.4	Effects on follow-up survey participation	90
3.4	Data & Methods.....	91
3.4.1	Questionnaire.....	94
3.5	Analysis Plan	95
3.5.1	Unit and item response	95
3.5.2	Nonresponse bias.....	96
3.5.3	Substantive responses.....	97
3.5.4	Effects on follow-up survey participation	97
3.6	Results.....	98
3.6.1	Unit and item response	98
3.6.2	Nonresponse bias.....	100

3.6.3	Substantive responses.....	104
3.6.4	Effects on follow-up survey participation	105
3.7	Conclusion	108
	References	112
	Appendix	117
4	Effects of Incentives in Smartphone Data Collection	119
4.1	Abstract.....	119
4.2	Introduction.....	119
4.3	The Influence of Incentives on Participation	120
4.4	IAB-SMART study design	124
4.4.1	Sampling frame and sample restrictions.....	125
4.4.2	Invitation and data request.....	127
4.4.3	Experimental design for incentive study	130
4.4.4	Analysis plan	132
4.5	Results.....	133
4.5.1	App installation	133
4.5.2	Number of initially activated data sharing functions.....	136
4.5.3	Deactivating functions.....	137
4.5.4	Retention	139
4.5.5	Analysis of costs.....	141
4.6	Summary.....	143
4.6.1	Limitations and future research	145
	References	148
	Appendix	153

5	Using Geofences to Collect Survey Data: Lessons Learned From the IAB- SMART Study	158
5.1	Abstract.....	158
5.2	Introduction.....	158
5.3	Design.....	160
5.4	Results.....	165
5.4.1	Number of triggered surveys and responses.....	166
5.4.2	Challenges	168
5.5	Conclusion - Lessons learned	174
5.6	Use of geofences in future research.....	177
	References	179
	Appendix	182
6	Conclusion	187

List of Figures

Figure 2.1: Index page for the web survey in the Digitalization version	50
Figure 2.2: Experimental Assignment and Response Rates (RR).....	52
Figure 2.3: Linear prediction of the estimated median time to complete the questionnaire in minutes for the Minimum Wage questionnaire (bars show 95% confidence intervals)	64
Figure 2.4: Predicted probabilities from the ordinal logistic regression model for perceived time and perceived burden in the Minimum Wage version by mode group (bars show 95% confidence intervals).....	64
Figure 2.5: Linear prediction of the estimated median time in minutes to complete the questionnaire for Digitalization questionnaire (bars show 95% confidence intervals)	65
Figure 2.6: Predicted probabilities from the ordinal logistic regression model for perceived time and perceived burden in the Digitalization version by mode group (bars show 95% confidence intervals).....	65
Figure 3.1 Overall design of the nutrition study	92
Figure 3.2: Proportion (points) of food entries for the nutrition question by day of the week Sunday (S) – Thursday (T) with 95% confidence intervals (lines).	104
Figure 4.1 Sample size at each stage of the IAB-SMART selection process	126
Figure 4.2: Screen-shots showing the five data sharing functions (top) and in-app Amazon.de vouchers (bottom). The app was programmed and offered in German, but direct translations into English are provided in the Figure.	129
Figure 4.3: Crossed experimental design with maximum incentive amounts for a six-month data collection period (N=2,853).....	131

Figure 4.4: Percentage of app installations of invited individuals, with 95% confidence intervals (N = 2,853), by incentive condition, maximum amount of incentives and welfare status	134
Figure 4.5: Percentage of app installations of invited individuals, with 95% confidence intervals (N = 2,853), by experimental groups and maximum amount of incentives, by welfare status	135
Figure 4.6: Mean number of initially activated data sharing functions, with 95% confidence intervals, conditional on installation, by incentive condition and maximum amount of incentives (N = 420).....	137
Figure 4.7: Percentages of participants who deactivated their function settings at least once, with 95% confidence intervals, conditional on installation, by incentive conditions, maximum amount of incentives and welfare status (N = 420)	139
Figure 4.8: Average time participants stayed in field, with 95% confidence intervals, i.e., time between first installation and deinstallation in percent, by incentive conditions and maximum amount of incentives (N = 420)	140
Figure 4.9: Average percent of points redeemed by participants, with 95% confidence intervals, by incentive conditions, maximum amount of incentives and welfare status (N=420)	142
Figure 4.10: Mean number of initially activated data-sharing functions with 95% confidence intervals, by incentive conditions and maximum amount of incentive, by welfare status (N = 420).....	153
Figure 4.11: Average time participants stayed in field with 95% confidence intervals, i.e., time between first installation and deinstallation in percent, by incentive conditions and maximum amount of incentive, by welfare status (N = 420) ..	154
Figure 4.12: Average percent of points redeemed by participants with 95% confidence intervals by incentive conditions and maximum amount of incentive, by welfare status (N=420)	155

Figure 4.13: Original (German) voucher flyer; experimental conditions are marked in red.	156
Figure 4.14: Voucher flyer (English translation); experimental conditions are marked in red.	157
Figure 5.1: The three events (Enter, Dwell and Exit) that the Google Geofence API measures (source: https://developers.google.com/location-context/geofencing/ , accessed: January 12, 2020).	164
Figure 5.2: Verification question that appeared as the first question upon accessing the geofence survey with the translation on the left.	165
Figure 5.3: Implemented job center geofences in Germany (N=410).....	167
Figure 5.4: Example plot for the distances between the job center and the custom function geolocation measures on the day of the geofence triggered survey.....	172
Figure 5.5: Example for distances between the job center and the custom function geolocation measures on the day of the geofence triggered survey.....	174
Figure 5.6: Screenshots and English translation of the geofence survey for IAB-SMART participants that received the survey invitation and answered the first question with “Yes”.	186

List of Tables

Table 1.1: List of each thesis paper with the corresponding project.	19
Table 2.1: Summary statistics for the time to complete the questionnaire in minutes by questionnaire version and mode group	57
Table 2.2: Summary of the six models for evaluating the response burden	58
Table 2.3: Proportions of perceived time and burden by questionnaire version.	60
Table 2.4: Wording and response options for the response burden indicators.	80
Table 2.5: Joint χ^2 values from margin contrast for Minimum Wage and Digitalization questionnaire versions and hypotheses 1-3.	81
Table 3.1: Summary of response rate in text message surveys.	84
Table 3.2: Experimental design of text message survey Wave 2 for question (Q) 1-8.	94
Table 3.3: Wording for questions 4 to 8 in the single-sitting and the modular design	95
Table 3.4: Number and percent of Respondents for total, 8 questions, 1-7 questions and each question by experimental design.	99
Table 3.5: Comparing nonresponse bias for single-sitting and modular text message survey design by comparing invited with respondent sample for each design	102
Table 3.6: Average marginal effects (AME) for logistic regression models to evaluate the impact of participation on follow-up surveys	106
Table 3.7: Questionnaire Wording for text message surveys (wave 1, wave 2 single-sitting and wave 2 modular)	117
Table 5.1: Number of IAB-SMART participants by the number of triggered surveys.	167
Table 5.2: Number of triggered surveys by the number of job centers.	168

1 Introduction

“The survey method has strengths and deficits that are reflections of the society that it measures; the very act of speaking candidly to a stranger is governed by norms that can and do change. Survey research has always and must always adapt to those changes.”

- Robert Groves (2011, p. 870)

Many areas, such as the government, market research, opinion research, social science, etc., and professions rely on the survey method to generate insights for their respective fields. For instance, governments use establishment surveys to monitor the economy, help managers make decisions and enable politicians to craft informed policies (Jones et al. 2013).

According to Robert Groves (2011), survey research is currently in its third era (1990 – present) since the field was established in the 1930s. The third era is characterized by technological advancements that can be grouped into two developments worth mentioning (see Groves 2011, p. 865). First, traditional offline survey modes such as face-to-face and telephone surveys have been mostly replaced by much cheaper web surveys (Callegaro and Yang 2018). Second, there is an increasing availability of organic data (also known as big data, sensor data, digital trace data, and social media data) that are data collected for a purpose other than research (e.g., Salganik 2017, Baker 2017). Both developments were long seen as competitors, with the survey profession fearing being replaced by organic data solutions (Callegaro and Yang 2018, Couper 2017). However, surveys are not being replaced but rather have shifted to web-based administration, and researchers worldwide are starting to evaluate the possibilities of combining so-called designed and organic data (Callegaro and Yang 2018).

As a member of the statistical methods department (KEM) at the Institute for Employment Research (IAB) and the Collaborative Research Center SFB 884 (SFB), my career as a survey methodologist started in 2015 with improving web data collection in establishment surveys and proceeded with novel approaches to mobile phone data collection, including mobile phone sensor data. Over the last decade, novel data collection approaches have increased in number due to the increased use of web surveys, the possibility of collecting a large variety of organic data, the combination of surveys and organic data and the overall development of information and communication technologies (ICTs). Past learnings of the survey profession are helpful to developing new design solutions for those novel data collection methods. However, the effects of new data collection design possibilities have yet to be explored. Understanding how design decisions affect the data generation process is crucial for assessing the validity and reliability of measurements and the explanatory power when inferring from a sample to a population (Kreuter and Peng 2014). Using data from three different projects, my thesis contains four studies focusing on the effects of novel designs on outcomes related to response burden and data quality.

The first study focuses on the difference in the response burden between the paper and web modes in an establishment survey. The second study evaluates how different administrations of a text message survey affect the response rates, nonresponse bias, substantive responses and the propensity to participate in a follow-up survey. The third study evaluates the effects of different incentive strategies on installation rates, activating data collection functions, withdrawing data collection consent and retention in a smartphone data collection project. The last study evaluates a novel technique called geofencing, that is, using geolocation data to trigger survey invitations over smartphones.

In the following, I briefly describe recent developments in ICT that led to a rethinking of how to conduct survey data collection similar to the telephone mode replacing personal interviews in the 1970s (Dillmann et al. 2010). To accomplish this, I use the web mode as the starting point. Furthermore, I provide a brief description of each project related to the submitted thesis papers.

In the beginning of the third era, as internet penetration was low, web surveys were used to complement other modes (Groves 2011). With increasing internet penetration rates worldwide, the web mode has become increasingly more reliable and works increasingly more as a standalone mode in many use cases, making web surveys one of the main data collection tools for market and social research (see, e.g., Keusch 2015). One reason for the web mode's increasing popularity may be the rise of do-it-yourself web survey tools (e.g., Survey Monkey or LimeSurvey) and in-house survey tools (see Callegaro and Yang 2018). Those web survey tools have high levels of usability, which makes them convenient for individuals to craft their own web survey without having a background in programming or survey methodology.

As web surveys enable timely and cost-effective data collection, web surveys have developed into one of the most important tools for professions requiring fast and cheap data, e.g., journalists and politicians. However, the survey market is full of statistics from web surveys using a nonprobability approach and ignoring lessons from the first era of survey research that relied on probability samples (Groves 2011). Survey professionals using the Total Survey Error (TSE, see Groves et al. 2011) framework as the main school of thought or guiding method to evaluate the quality of surveys or data in general are challenged in competing with cheap and timely nonprobability samples by using well-maintained sample frames to invite individuals based on a probability sample. A good example that overcomes this challenge is the German internet panel (GIP), which uses a probability sample

based on the general population in Germany and recruits individuals face-to-face to minimize the nonresponse bias associated with the web survey (Blom et al. 2015).

Web surveys may have become easier to use; however, designing a web survey may have special requirements in certain settings, such as when surveying establishment populations. Independent of the data collection mode, questions in personal surveys address opinions, attitudes and/or behaviors. Establishment surveys may ask about different areas of the establishment, e.g., human resources, business finances, etc. This knowledge may be distributed over different employees. Therefore, the web survey must be shareable with colleagues who own specific knowledge, and the survey requires a navigation tool indicating which sections of the questionnaire may need which expertise.

A major concern in establishment surveys is response burden, that is, the strain experienced by respondents when they respond to a survey. Each survey response uses precious time of an establishment that could have been used to generate a product or service (e.g., Haraldsen et al. 2013). Therefore, response burden is regarded as an important issue in the establishment survey context, and researchers aim to reduce the response burden for establishments as much as possible (see, e.g., Giesen et al. 2018). Using a web mode in an establishment survey may be a more cost-effective mode of choice but may also have an effect on response burden. The first paper of my thesis evaluates how using a web mode in an establishment survey affects response burden compared to using a paper mode (see chapter 2).

At the beginning of the third era, web surveys were usually designed for a PC or laptop screen. These days, individuals access the internet using mobile devices such as smartphones and tablets. With the increased use of mobile devices, researchers must design web surveys that are compatible with a smaller screen size and optimize web surveys for smartphones (Toepoel et al. 2020). Therefore, web surveys can be considered mixed-

device surveys, and survey designers are challenged to design online surveys that are suitable for larger (e.g., laptops or PCs) or smaller (e.g., smartphones) screens (Toepoel and Lugtig 2015).

Groves's statement that the third era is characterized by the rise of web surveys may be true for first-world countries with high internet penetration rates but not for regions where internet penetration rates are low. In 2014, PEW published a report on mobile and internet use for emerging nations (PEW 2014). The report shows that the mobile phone penetration rates lie between 55 and 95%, while the internet penetration level is lower in each of the listed countries. Therefore, web surveys may not be used for all populations. It may make more sense to use other modes such as text message surveys, which is a mode that has resulted from ICT development over the last decade; however, there is currently little experience available regarding how to best apply this method. The second paper of my thesis assesses different strategies of using a text message survey in Egypt concerning differences between response rates, completeness, non-response bias, substantive responses and the effect on response rates in follow-up surveys (see chapter 3).

To contrast the term designed data, essentially meaning survey data, Groves (2011) uses the term organic data. Organic data can be seen as another term for big data, which in itself has a wide range of definitions (see Dutcher 2014). Researchers seek to grasp, define, distinguish and summarize the different terms and types of big or organic data. Couper (2017), for instance, summarizes the different types and terms as follows:

“[...] (a) administrative data, which are provided by persons or organizations for the administration of a program (e.g., electronic medical records, insurance records, bank records, tax records, registers); (b) transaction data, which are generated as an automatic by-product of financial or other transactions and activities (e.g., credit card transactions, online purchases); (c) sensor data (e.g., satellite imaging, road sensors, climate sensors); (d) tracking device data (e.g., GPS, mobile phones); (e) behavioral data (e.g., online searches, page views, cookie data); and (f) social media data, which are created by people with the express purpose of sharing with (at least some) others (e.g., Facebook, Twitter). The focus here is

primarily on the last two types, sometime referred to as big social data (Lampe et al. 2014), organic data (Groves 2011), found data (AAPOR 2015, Schober et al. 2016), naturally occurring data (Gelman et al. 2014), or data in the wild (Ang et al. 2013).”

Couper 2017, p. 12.14

For my thesis, it is important to acknowledge that the data collection process for organic data is not designed for research but for another purpose (Salganik 2018). Organic data are intriguing for researchers because they solve problems for the survey profession, such as increased costs for conducting surveys and decreased response rates, survey biases and response burden (Amaya et al. 2020).

Organic data contain information that may bring valuable insights for research and other purposes. For instance, the Institute for Employment Research (IAB) conducts the Integrated Employment Biography (IEB). The IEB consists of notifications of employers regarding their employees and employment agencies regarding their customers sent to the Federal Employment Agency (BA) in Germany. One of the main purposes of collecting these data is not to use it for research but rather to calculate several dues and claims, e.g., dues for unemployment insurance and claims in the case of unemployment. This information can be used to create a dataset containing employment biographies of labor market participants all the way back until 1975, e.g., information about their periods of employment, unemployment, welfare reciprocity and other topics (for more details, see Antoni et al. 2019).

Sometimes the volume of organic data is confused with information (Groves 2011); therefore, organic data users are prone to the fallacy that organic data are error free. However, organic data are not error free and sometime face challenges and confidentiality concerns (Amaya et al. 2020). The survey profession spent decades defining and assessing different sources of error when using the survey method. Many of these sources of error can be

applied to organic data. For instance, using IEBs requires researchers to account for limitations in coverage, as some occupational groups, such as self-employed and civil servants, are not represented in the IEB (e.g., Antoni et al. 2019). Therefore, the IEB cannot make any statements about those occupational groups and suffers from coverage bias when making inferences regarding the whole German labor market population. Furthermore, if labor market participants become self-employed or civil servants, gaps within the data appear (missing data errors) that may be interpreted as the labor market participant leaving the labor market instead of transferring to another profession that is not captured by the data (e.g., Antoni et al. 2019). Additionally, some variables are not needed for the intended purpose to calculate dues and claims (e.g., educational status) and therefore suffer from measurement errors (e.g., Fitzenberger et al. 2005). As a result, employers pay less attention to reporting recent educational status data, e.g., if an employee earns an additional degree. Therefore, some variables in the IEB suffer from measurement errors that require adjustments, such as imputation.

Designed and organic data contain sources of error along the themes of the well-known TSE framework, such as validity, missingness and representativeness (Baker 2017). Organic data, however, also introduce new sources of error. Research using Twitter data, for instance, suffers from query errors (Hsieh and Murphy 2017), which occur when specifying the keywords used for searching Twitter posts.

The fact that designed and organic data contain different error sources does not mean that they should not be used. Understanding these sources of error helps researchers and users understand the validity of their analyses and findings, identify which data best fit the research question, predict which type of data suffers from which sources of error, and account for or eliminate these sources of error beforehand (e.g., Amaya et al. 2020, Kirchner et al. 2020, Kreuter and Peng 2014).

The jewel in the crown combines designed and organic data and employs their strengths to complement each other (e.g., Couper 2017, Lazer and Radford 2017). Survey data are suitable for collecting opinions and attitudes (the why), while organic data are suitable for collecting behavioral data (the what) (Callegaro and Yang 2018). Combining both types of data enables researchers to add more context to their designed or organic data and may be helpful to evaluate and harmonize potential sources of error in both data types (Link et al. 2014, Kirchner et al. 2020).

Combining designed and organic data or integrating organic data collection processes into a designed data collection process is one of the recent major challenges. One way to combine designed and organic data collection is to use smartphone apps that collect both survey and smartphone sensor data. My third and fourth thesis papers draw on the IAB-SMART project that uses such a data collection approach. While the third thesis paper evaluates the effects of different incentive strategies on installation rates and retention (see chapter 4), the fourth thesis paper evaluates a novel approach called geofencing, that is, using geolocation data to trigger survey invitations (see chapter 5).

I briefly discuss the developments of web surveys, the rise of mobile phone surveys, the increasing significance of organic data for the survey profession and the need to combine designed and organic data. Against this background, the survey profession is facing new unexplored possibilities to design and modernize data collection approaches resulting from recent developments of ICTs that changed the survey landscape. My thesis explores the potential of design and modernized data collection methods by evaluating different research designs in each of my submitted papers.

My thesis consists of four papers from three projects (see Table 1.1). The remaining part of the introduction discusses each project and briefly describes the content of each paper. First, I describe the “Quality in Establishment Surveys (QuEst)” project, which set out to

evaluate methodological differences between the paper and web modes in establishment surveys. Second, I describe a project in Egypt in which we tested how to conduct a short text message survey. Third, I describe the IAB-SMART study, which is an app research project that collected various smartphone sensor and survey data. For my thesis, I submit two papers from the IAB-SMART study. The first paper evaluates the effect of different incentive strategies on installation rates, activating data collection, withdrawing data collection and retention in an experimental setting. The second paper evaluates a novel technique to conduct survey data collection dependent on geolocation data collection from the Google Geofence API.

Table 1.1: List of each thesis paper with the corresponding project.

Paper	Project	Written with
Comparing Response Burden Between Paper and Web Modes in Establishment Surveys	Quality in Establishment Surveys (QuEst)	Stephanie Eckman and Ruben Bach
Comparing Single-sitting Versus Modular Text Message Surveys in Egypt	Project A9: Survey mode, survey technology and technology innovations in data collection of the Deutsche Forschungsgesellschaft (DFG) funded project: SFB 884 "Political Economy of Reforms"	Florian Keusch and Markus Frölich
Effects of Incentives in Smartphone Data Collection.	IAB -SMART	Frauke Kreuter, Florian Keusch, Mark Trappmann and Sebastian Bähr
Using Geofences to Collect Survey Data: Lessons Learned From the IAB-SMART Study	IAB -SMART	Mark Trappmann, Florian Keusch, Sebastian Bähr and Frauke Kreuter

1.1 Quality in Establishment Surveys (QuEst)

The project was initiated to prepare the IAB establishment survey to move their data collection from paper to the web. The web mode offers several advantages over paper regarding data collection costs. For instance, web surveys can eliminate mailing costs by inviting establishments to participate via e-mail. If an e-mail address is not available, web surveys still reduce mailing costs by sending mail invitations containing only an invitation letter rather than a large paper questionnaire and return envelope. Furthermore, as respondents digitalize their answer themselves during the response process, web surveys reduce data entry costs.

The web mode also offers advantages regarding data quality. Web questionnaires can provide feedback to respondents and therefore may increase data quality (Couper 2008, Conrad et al. 2007). If respondents submit an unlikely answer, plausibility checks can ask respondents to re-evaluate their answers, which could reduce the need for data editing. Furthermore, researchers can offer definitions and additional information about questions. Future web surveys may even include chatbots that can address respondents' questions during the response process (Lagerstøm 2018). Additionally, the web mode can manage calculation and counting tasks to simplify responses (Giesen et al. 2009, Giesen 2007). Especially in establishment surveys, which often require responses from multiple respondents, web surveys may simplify the response process within the establishment as respondents can distribute a link for a web survey easily via email while a paper questionnaire is more cumbersome to distribute to multiple respondents. Finally, web surveys enable a complex filter and skip pattern design and thus only show applicable items to each respondent.

In the establishment survey context, little is known about the effects of the web mode. By implementing an experimental setting with five different mode conditions (*Paper-only*,

Web-only, Choice, Paper-followed by Choice, Web followed by Choice) the QuEst project helped to narrow the gap.

All five mode conditions received a postal invitation letter to the study. Overall, one invitation letter and two reminders were sent. However, the forms and choices varied over the mode groups. For the *Paper-only* group, we mailed establishments a cover letter with information about the study and a paper questionnaire. Respondents could only respond by sending back a completed questionnaire. For the *Web-only* group, establishments received a cover letter with a link and a request to complete our online questionnaire. The establishments in the *Choice* groups received a cover letter and the same paper questionnaire as that provided to the *Paper-only* group. The cover letter offered a web link and presented the respondents with the option to choose between the paper and web modes. For the paper followed by choice mode, respondents received the same invitation letter and reminder materials as those used in the *Paper-only* group. The second reminder contained the materials as those provided to the *Choice* group. For the web followed by paper mode, respondents received the same invitation letter and reminder materials as those provided to the *Web-only* group. The second reminder contained the materials as those provided to the *Choice* group.

In the first step, the mode experiment was analyzed considering the response rate, composition effects and item nonresponse rate (see Haas et al. 2016). Concerning the response rates, the authors found that web performs poorly compared to paper (5.6% vs. 11.7%). However, offering both modes in all mailings performs as well as the paper mode. This outcome is interesting as it contradicts well-established results that providing respondents with a mode choice lowers response rates (see Medway and Fulton 2012). In terms of

composition effects, the authors found no systematic differences between mode conditions. Regarding item response rates, the results suggest that web respondents have lower item nonresponse rates than paper respondents.

The paper for this thesis evaluates how much the web mode affects the response burden during the data collection process compared to the paper mode. Response burden is a multifaceted concept influenced by motivation, task difficulty, survey effort and respondent perception (Yan et al. 2020). Especially in establishment surveys, data collectors should monitor and reduce burden to the fullest possible extent (e.g., see European Commission 2011). Giesen et al. (2018) even state that burden management has to become a key element for all national statistical institutes.

Modernizing establishment survey designs and moving the data collection for establishments towards the web mode may help data collectors reduce establishments' response burden. However, previous research is inconclusive regarding the effects of switching modes to the web on response burden. The first paper of my thesis, "*Comparing Response Burden Between Paper and Web Modes in Establishment Surveys*" (Haas et al. forthcoming), evaluates the differences in the response burden between paper and the web. The paper compares how the actual and perceived response burdens differ when respondents complete a survey in the paper mode, the web mode or are allowed to choose between the two modes. The results show that in the web mode respondents estimate the time to complete the questionnaire to be lower than paper respondents while there are no differences between paper and the web in the perceived response time and perceived burden. No evidence showing that researchers should be concerned that the response burden is increased when conducting an establishment survey using a web survey instead of a paper survey is found.

1.2 Project A9: Survey mode, survey technology and technology innovations in data collection

Project A9 is embedded in a larger research agenda called Sonderforschungsbereich 884 (SFB 884), which is funded by the Deutsche Forschungsgemeinschaft (DFG). SFB 884 is an interdisciplinary research group consisting of an interdisciplinary workforce of researchers from economics, political science, sociology, statistics and computer science. It aims to produce scientific insights into the successes and failures of reforms that address economic and social challenges. Project A9 works closely together with other projects from SFB 884 to further develop data collection methods in terms of improved data quality and lower costs of future projects. The project does focus on several developments and technological innovations in survey research. For this thesis, however, the focus is on mobile phone data collection.

Part of the research agenda of project A9 is the exploration and evaluation of new data collection methods in certain research areas. Examples include (1) avoiding interviewee fatigue with the use of ultra-short but high-frequency surveys, (2) combining surveys with information treatments and (3) interviewing difficult-to-interview populations. These three research areas are addressed in the paper selected for my thesis: “*Comparing Single-sitting Versus Modular Text Message Surveys in Egypt*”.

The experiment described in the thesis paper piggybacks on a larger economic field study evaluating the effects of information treatments on nutritional health of kindergarten children in Egypt. As internet penetration rates (43%, see PEW 2014) and landline penetration rates (29%, see PEW 2015) are low in Egypt, web and telephone surveys suffer from undercoverage. However, Egypt has a rather high mobile phone penetration rate of 88% (see PEW 2014). Therefore, a text message survey was conducted to measure differences in behavior due to an information treatment. To date, little is known about how to best

administer text message surveys efficiently. Therefore, two different designs of automated text message surveys were experimentally compared. In the first design, *single-sitting*, respondents automatically receive a text message with a new question once they replied to a question. In the second design, *modular*, respondents received a new question each day regardless of whether they had responded to the previous question. Overall, 1,081 Egyptian parents of kindergarten children who own a mobile phone were invited to participate in a text message survey with eight questions on the nutritional behavior of their children. The sample was randomly assigned to one of the experimental groups. In short, compared to the *single-sitting* group, the *modular* group achieved a higher number of answered questions but had fewer fully completed questionnaires. In addition, the experimental groups differed regarding substantive responses to behavioral questions. No differences concerning nonresponse bias or the probability of responding to a follow-up survey were found.

1.3 IAB-SMART

In 2018, we (Frauke Kreuter, Florian Keusch, Mark Trappmann, Sebastian Bähr and myself) conducted the IAB-SMART app study that collected survey data and smartphone sensor data. The data included geolocations, telephone and text message logs, characteristics of social networks from smartphone contacts, mobility data and smartphone usage data. The IAB-SMART study had the intention of exploring the feasibility of how the combination of passively collected smartphone and survey data can be helpful to replicate old insights and gain new insights into labor market research. One of the main goals was to replicate the results from one of the first and most famous studies that applied a combination of various methods, namely, the Marienthal study by Jahoda et al. (1939).

Marienthal was a small city near Vienna that was built around a textile factory. Almost everybody living in Marienthal was employed in this factory. However, due to the economic crisis in 1929, the factory was closed in 1930, and all workers of the factory were unemployed. What was a tragedy for the community of Marienthal was an opportunity for social scientists including Marie Jahoda and Paul Lazarsfeld to research the effects of unemployment on the social life in a community. The researchers used a variety of methods, such as qualitative interviews, field observations in the streets, stocktaking of household belongings, etc.

One of their most prominent findings was that unemployment results in a loss of structure in the daily lives of citizens. When Marienthal's citizens were employed, they valued their time in a remarkably high manner due to its scarcity. With the loss of employment, time was abundant in daily living, which resulted in less reading, canceling club memberships and a slower walking pace. Interestingly, as women also had to manage household affairs (i.e., cooking, cleaning and caring for children), they had a purpose that structured their daily lives and were less affected by the effects of unemployment compared to men.

The Marienthal study was possible with the help of a large research team that performed timely expensive data collection. Due to the ubiquity of smartphones in everyone's daily lives, the results from the Marienthal study may be replicated with less manpower and in a more cost-efficient way. Smartphones can collect various data, such as geolocation data, mobility data and app usage data. As individuals use smartphones in their daily lives, smartphone data may inform researchers about smartphone owner behavior and may be used to generate insights for social science.

In the past few years, many researchers have started to evaluate the use of smartphone or mobile phone data for their research projects (e.g., Ben-Zeev et al. 2015, Elevelt et al.

2019, Goodspeed et al. 2018, Harari et al., 2017, Lathia et al. 2017, MacKerron and Mourato 2013, Montag et al. 2015, Smeets et al. 2019, Scherpenzel 2017, Sugie 2018, Wang et al. 2014, York Cornwell and Cagney, 2017, Katevas et al. 2018). Most of those studies, however, are based on convenience samples, have a small sample size (less than 100 participants), have a short data collection period (less than a month) and/or are limited in evaluating potential sources of error in their data. The IAB-SMART study overcame these limitations and is a perfect example of the role of the survey profession in exploring new opportunities and challenges to design new data collection approaches and integrate combined survey and organic data while assessing different sources of error along the themes of the total survey error framework (TSE).

The main reason why the IAB-SMART study enabled researchers to assess sources of error around the TSE framework is that study participants were selected from the panel study “Labour Market and Social Security” (PASS) (Trappmann et al. 2019). PASS is a household survey of the general population in Germany with annual waves of data collection. The focus of PASS is on unemployment, poverty and reciprocity of state transfers. Using PASS as a starting point to invite study participants has the benefit that a relationship of trust between study participants and the IAB is already implemented. Furthermore, linking PASS and IAB-SMART data provides background information about participants and nonparticipants, making it possible to assess coverage and nonresponse bias in the IAB-SMART participant sample (see Keusch 2020, Keusch *under review*).

Overall, 685 of the 4,293 invited PASS respondents installed the app. Recruiting participants to install a research app that collects various abstract types of data is one of the new design challenges mentioned above. Installing an app and keeping the app installed is quite different from responding to a one-time web survey. Furthermore, when collecting passive mobile phone data or linking app data to other data sources (e.g., PASS panel

data), researchers have to consider further data protection measures and communicate those in a simple, transparent and comprehensive way. This gets harder as more different types of data are collected. To meet ethical and legal standards, we implemented the principles of the General Data Protection Regulation (GDPR) on informed consent. We developed a comprehensive strategy to communicate the purpose of our study to invited individuals and sought to lower their privacy concerns as much as possible so that they would agree to our data collection (see Kreuter et al. 2020 for more details).

It is quite common for most smartphone owners to install an app and to accept that app's permission to obtain access to all types of smartphone data. However, being invited to install an app that collects several different types of smartphone data may be suspicious. One could argue that apps such as WhatsApp and Facebook also collect many types of behavioral data and that it should not be a problem for most individuals to give their data to unknown researchers. However, installing the Facebook app follows the intention of using a service that enables smartphone owners to connect with their families and friends or to entertain themselves. A research app can hardly make a similar offer and has to appeal to other factors to motivate individuals to participate, e.g., using convincing arguments that data are used for the social good.

A strategy to motivate individuals to participate in a research project or survey is using incentives. In surveys, monetary incentives are known to compensate respondents and increase response rates (e.g., James and Bolstein 1990, Church 1993, Willimack et al. 1995, Singer et al. 1999, Singer 2002, Toepoel 2012, Pforr 2016). The rule of thumb is that cash incentives are more effective than gifts or lotteries and that unconditional (also known as (aka) prepaid) incentives are more effective than conditional (aka promised) incentives. (Singer and Ye 2013). However, little is known about how incentives work in studies involving smartphone sensor data. Smartphone data may be perceived as more

valuable than survey data, and it may be harder to recruit participants for an app study involving smartphone sensor data than for regular surveys. Furthermore, the IAB-SMART study was set to a field period of 180 days. It was important that the app collects daily data from each individual without any participant-mediated gaps. The third paper for my thesis, *“Effects of Incentives in Smartphone Data Collection”* (Haas et al. 2020a), addresses this gap. It addresses the use of incentives in smartphone data collection and evaluates the effects of different incentive strategies on downloading the app, activating data collection, withdrawing data collection and retention in an experimental setting.

The IAB-SMART app contained the Google Geofence API, which allowed us to administer surveys that were triggered by geographical areas. This survey data collection technique is called geofencing. Against the background of survey data collection, a geofence can be described as a geographical area that triggers a survey by entering this area, dwelling within this area for a defined amount of time and/or exiting this area. Therefore, geofencing combines organic and survey data in an interesting way by using real-time organic data, in our case geolocation data, to trigger a survey as a novel design feature to recruit respondents to respond to a survey.

Geofences are already used in other contexts (e.g., marketing and retail) but are still underutilized in social research. The fourth paper, *“Using Geofences to Collect Survey Data: Lessons Learned From the IAB-SMART Study”* (Haas et al. 2020b), addresses the challenge of conducting survey data collection that is dependent on geolocation data collection from the Google Geofence API. We implemented a geofence survey and geofenced 410 job centers with the Google Geofence API. Overall, the app sent 230 geofence-triggered survey invitations to 107 participants and received 224 responses from 104 participants. This paper shows how the geofence method was applied within the IAB-SMART project and provides the reader with six lessons researchers should consider

when designing a geofence data collection project to increase data quality and minimize costs.

1.4 The modernization of data collection methods

I have briefly presented each project from which my submitted thesis papers resulted. Traditionally, dissertations focus on one topic that is wholesomely and thoroughly examined. In this regard, the submitted papers of my thesis may be too diversified over different topics and approaches. Over the last decade, however, the survey profession has also become much more diversified. Increasingly more frequently, survey methodologists and statisticians find themselves caught between using survey data collection and other various types of data collection and combinations thereof that belong to other professions. My thesis mirrors this development and the need for the diversification of the field by not specializing on one topic but rather by assessing different topics that address new challenges in survey design.

References

- Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, 8(1), 89–119. <https://doi.org/10.1093/jssam/smz056>.
- Antoni, M., Ganzer, A., and vom Berge, P. (2019). *Sample of Integrated Labour Market Biographies Regional File (SIAB-R) 1975 - 2017*. <https://doi.org/10.5164/IAB.FDZD.1904.EN.V1>.
- Baker, R. (2017). Big Data. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, . . . B. T. West (Eds.), *Total Survey Error in Practice* (pp. 47–69). Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781119041702.ch3>.
- Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., and Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3), 218–226. <https://doi.org/10.1037/prj0000130>.
- Blom, A. G., Gathmann, C., and Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population. *Field Methods*, 27(4), 391–408. <https://doi.org/10.1177/1525822X15574494>.
- Callegaro, M., and Yang, Y. (2018). The Role of Surveys in the Era of “Big Data”. In D. L. Vannette and J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 175–192). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_23.
- Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly*, 57(1), 62. <https://doi.org/10.1086/269355>.

- Conrad, F. G., Schober, M. F., and Coiner, T. (2007). Bringing features of human dialogue to web surveys. *Applied Cognitive Psychology*, 21(2), 165–187.
<https://doi.org/10.1002/acp.1335>.
- Couper, M. P. (2017). New Developments in Survey Data Collection. *Annual Review of Sociology*, 43(1), 121–145. <https://doi.org/10.1146/annurev-soc-060116-053613>
- Couper, M. P. (2008). *Designing Effective Web Surveys*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511499371>.
- Dillman, D. A., Reips, U.-D., and Matzat, U. (2010). Advice in surveying the general public over the internet. *International Journal of Internet Science*, 5(1), 1–4. Retrieved from <http://kops.uni-konstanz.de/handle/123456789/28669>.
- Dutcher, J. (2014). What Is Big Data? – Blog. Retrieved from <https://datascience.berkeley.edu/what-is-big-data/>.
- Elevelt, A., Lugtig, P., and Toepoel, V. (2019). Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process? Advance online publication.
<https://doi.org/10.18148/SRM/2019.V13I2.7385>
- European Commission (2011). *European Statistics Code of Practice for the National and Community Statistical Authorities*. Adopted by the European Statistical System Committee, September 28, 2011. Retrieved from <https://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15>.
- Fitzenberger, B., Osikominu, A., and Völter, R. (2006). Imputation rules to improve the education variable in the IAB employment subsample. *Schmollers Jahrbuch: Journal of Contextual Economics*, 126(3), 405–436.

- Giesen, D., Vella, M., Brady, C. F., Brown, P., Ravindra, D., and Vaasen-Otten, A. (2018). Response Burden Management for Establishment Surveys at Four National Statistical Institutes. *Journal of Official Statistics*, 34(2), 397–418.
<https://doi.org/10.2478/jos-2018-0018>
- Giesen, D., Morren, M., and Snijkeres, G. (2009). The effect of survey redesign on response burden: An evaluation of the redesign of the SBS questionnaires. In *3rd European Survey Research Association Conference (ESRA)*. Symposium conducted at the meeting of European Survey Research Association, Warsaw, Poland.
- Goodspeed, R., Yan, X., Hardy, J., Vydiswaran, V. G. V., Berrocal, V. J., Clarke, P., . . . Veinot, T. (2018). Comparing the Data Quality of Global Positioning System Devices and Mobile Phones for Assessing Relationships Between Place, Mobility, and Health: Field Study. *JMIR MHealth and UHealth*, 6(8), e168.
<https://doi.org/10.2196/mhealth.9771>
- Groves, R. M. (2011). Three Eras of Survey Research. *Public Opinion Quarterly*, 75(5), 861–871. <https://doi.org/10.1093/poq/nfr057>.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey Methodology* (2nd ed.). Hoboken: John Wiley & Sons. Retrieved from <http://gbv.eblib.com/patron/FullRecord.aspx?p=819140>.
- Haas, G.-C., Kreuter, F., Keusch, F., Trappmann, M. and Bähr, S. (2020a). Effects of Incentives in Smartphone Data Collection. In *Big Data Meets Survey Science* (eds C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japiec, A. Kirchner, S. Kolenikov and L.E. Lyberg). doi:10.1002/9781118976357.ch13.

- Haas, G.-C., Trappmann, M., Keusch, F., Bähr, S. and Kreuter, F. (2020b). *Using Geofences to Collect Survey Data: Lessons Learned From the IAB-SMART Study*. <https://doi.org/10.13094/SMIF-2020-00023>.
- Haas, G.-C., Eckman, S. and Bach, R. (forthcoming). Comparing Response Burden Between Paper and Web Modes in Establishment Surveys. *Journal of Official Statistics*.
- Haas, G.-C., Eckman, S., Bach, R., and Kreuter, F. Is Moving Establishment Surveys from Mail to Web a Good or Bad Decision in Terms of Performance and Data Quality? In *Proceedings of the International Conference for Establishment Surveys 2016 (ICES-V)*, Geneva, Switzerland. Retrieved from https://ww2.amstat.org/meetings/ices/2016/proceedings/ICESV_TOC.pdf.
- Haraldsen, G., Jones, J., Giesen, D., and Zhang, L.-C. (2013). Understanding and Coping with Response Burden. In G. Snijders, G. Haraldsen, J. Jones, and D. K. Willimack (Eds.), *Designing and Conducting Business Surveys* (pp. 219–252). Hoboken, New Jersey: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118447895.ch06>.
- Harari, G. M., Gosling, S. D., Wang, R., Chen, F., Chen, Z., and Campbell, A. T. (2017). Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior*, 67, 129–138. <https://doi.org/10.1016/j.chb.2016.10.027>
- Hsieh, Y. P., and Murphy, J. (2017). Total Twitter Error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, . . . B. T. West (Eds.), *Total Survey Error in Practice* (pp. 23–46). Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781119041702.ch2>.

- Jahoda, M., Lazarsfeld, P. F., Zeisel, H., and Fleck, C. (1939). *Marienthal: The sociology of an unemployed community*. New Brunswick, N.J.: Transaction.
- James, J. M., and Bolstein, R. (1990). The Effect of Monetary Incentives and Follow-Up Mailings on the Response Rate and Response Quality in Mail Surveys. *Public Opinion Quarterly*, 54(3), 346. <https://doi.org/10.1086/269211>.
- Jones, J., Snijkers, G., and Haraldsen, G. (2013). Surveys and Business Surveys. In G. Snijkers, G. Haraldsen, J. Jones, and D. K. Willimack (Eds.), *Designing and Conducting Business Surveys* (pp. 1–38). Hoboken, New Jersey: John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118447895.ch01>.
- Keusch, F., Bähr, S., Haas, G., Kreuter, F., Trappmann, M. (under review). Nonparticipation in Smartphone Data Collection Using Research Apps. *Journal of the Royal Statistical Society: Series A*.
- Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., and Trappmann, M. (2020). Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey. *Sociological Methods & Research*, 004912412091492. <https://doi.org/10.1177/0049124120914924>.
- Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, 65(3), 183–216. <https://doi.org/10.1007/s11301-014-0111-y>.
- Kirchner, A., Hochfellner, D., and Bender, S. (2017). Big data infrastructure at the Institute for Employment Research (IAB). In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West (Eds.), *Total survey error in practice* (pp. 458–465). John Wiley & Sons, Inc. Wiley Series in Survey Methodology <https://doi.org/10.1002/9781119041702.ch21>.

- Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., and Trappmann, M. (2020). Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent. *Social Science Computer Review*, 38(5), 533–549. <https://doi.org/10.1177/0894439318816389>.
- Kreuter, F. and Peng, R. D. (2014). Extracting Information from Big Data: Issues of Measurement, Inference and Linkage. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good* (pp. 257–275). New York: Cambridge University Press. <https://doi.org/10.1017/CBO9781107590205.016>.
- Lazer, D., and Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, 43(1), 19–39. <https://doi.org/10.1146/annurev-soc-060116-053457>.
- Lagerstøm, B. (2018). Chatbots as digital interviewers. *Paper presented at the International Household Nonresponse Workshop*, August 22–24, 2018. Budapest, Hungary.
- Lathia, N., Sandstrom, G. M., Mascolo, C., and Rentfrow, P. J. (2017). Happier People Live More Active Lives: Using Smartphones to Link Happiness and Physical Activity. *PloS One*, 12(1), e0160589. <https://doi.org/10.1371/journal.pone.0160589>.
- Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., and 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(4), 779–787. <https://doi.org/10.1093/poq/nfu054>.
- Medway, R. L., and Fulton, J. (2012). When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates. *Public Opinion Quarterly*, 76(4), 733–746. <https://doi.org/10.1093/poq/nfs047>.

- MacKerron, G., and Mourato, S. (2013). Happiness is greater in natural environments. *Global Environmental Change*, 23(5), 992–1000. <https://doi.org/10.1016/j.gloenvcha.2013.03.010>.
- Montag, C., Błazzkiewicz, K., Sariuska, R., Lachmann, B., Andone, I., Trendafilov, B., . . . Markowetz, A. (2015). Smartphone usage in the 21st century: Who is active on WhatsApp? *BMC Research Notes*, 8, 331. <https://doi.org/10.1186/s13104-015-1280-z>.
- Pew Research Center (2014). Emerging Nations Embrace Internet, Mobile Technology. Retrieved from <https://www.pewresearch.org/global/2014/02/13/emerging-nations-embrace-internet-mobile-technology/> (accessed February 2021).
- Pew Research Center (2015). Internet Seen as Positive Influence on Education but Negative on Morality in Emerging and Developing Nations. Retrieved from <https://www.pewresearch.org/global/2015/03/19/internet-seen-as-positive-influence-on-education-but-negative-influence-on-morality-in-emerging-and-developing-nations/> (accessed February 2021).
- Pforr, K. (2016). *Incentives*. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/GESIS-SG_EN_001.
- Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton, Oxford: Princeton University Press.
- Scherpenzeel, A. (2017). Mixing Online Panel Data Collection with Innovative Methods. In S. Eifler and F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 27–49). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15834-7_2.

- Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse*, pp. 163-177. John Wiley & Sons New York.
- Singer, E., and Ye, C. (2013). The Use and Effects of Incentives in Surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 112–141.
<https://doi.org/10.1177/0002716212458082>.
- Singer, E., R. M. Groves, and A. D. Corning (1999). Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly* 63 (2), 251-260.
- Smeets, L., Lugtig, P., and Schouten, B. (2019). Automatic Travel Mode Prediction in a National Travel Survey. *CBS Discussion Paper*. Retrieved from https://www.cbs.nl/-/media/_pdf/2019/51/dp%20smeets-lugtig-schouten%20-%20vervoermiddelpredictie.pdf.
- Sugie, N. F. (2018). Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups. *Sociological Methods & Research*, 47(3), 458–491.
<https://doi.org/10.1177/0049124115626176>
- Toepoel, V., Lugtig, P. and Schouten, B. (2020). Active and passive measurement in mobile surveys. *The Survey Statistician* 82, 14-26.
- Toepoel, V. and Lugtig, P. (2015). Online Surveys are Mixed-Device Surveys. Issues Associated with the Use of Different (Mobile) Devices in Web Surveys. Advance online publication. <https://doi.org/10.12758/MDA.2015.009>.
- Toepoel, V. (2012). Effects of Incentives in Surveys. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 209–223). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3876-2_13.

- Trappmann, M., Bähr, S., Beste, J., Eberl, A., Frodermann, C., Gundert, S., . . . Wenzig, C. (2019). Data Resource Profile: Panel Study Labour Market and Social Security (PASS). *International Journal of Epidemiology*, 48(5), 1411-1411g.
<https://doi.org/10.1093/ije/dyz041>
- Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., . . . Campbell, A. T. (09132014). StudentLife. In A. J. Brush, A. Friday, J. Kientz, J. Scott, and J. Song (Eds.), *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 3–14). New York, NY, USA: ACM.
<https://doi.org/10.1145/2632048.2632054>.
- Willimack, D. K., H. Schuman, B.-E. Pennell, and J. M. Lepkowski (1995). Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to- Face Survey. *Public Opinion Quarterly*, 59(1), 78-92.
- York Cornwell, E., and Cagney, K. A. (2017). Aging in Activity Space: Results From Smartphone-Based GPS-Tracking of Urban Seniors. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, 72(5), 864–875.
<https://doi.org/10.1093/geronb/gbx063>
- Yan, T., Fricker, S., and Tsai, S. (2020). Response Burden: What Is It and What Predicts It? In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, and A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 193–212). Wiley. <https://doi.org/10.1002/9781119263685.ch8>.

2 Comparing the Response Burden between Paper and Web Modes in Establishment Surveys

2.1 Abstract

Previous research is inconclusive regarding the effects of paper and web surveys on response burdens. We conducted an establishment survey with random assignment to paper and web modes to examine this issue. We compare how the actual and perceived response burdens differ when respondents complete a survey in the paper mode, in the web mode or when they are allowed to choose between the two modes. Our results show that in the web mode, respondents have a lower estimated time to complete the questionnaire, while we do not find differences between paper and the web on the perceived response time and perceived burden. Even though the response burden in the web mode is lower, our study finds no evidence of an increased response burden when moving an establishment survey from paper to the web.

2.2 Introduction

Data on establishments are essential for monitoring national and international economies, e.g., to help managers make decisions and to enable politicians to craft informed policies (Jones et al. 2013). A large proportion of establishment data originates from surveys. However, for most establishments, responding to a survey is a task unrelated to business production, which potentially takes employee time away from other essential tasks (Willimack and Nichols 2010). This article is particularly concerned with response burden, which unfortunately is loosely defined in the literature (Yan et al. 2020). We define response burden as the strain experienced by respondents while they respond to a survey. Factors affecting response burden are multifaceted and include questionnaire design, content and length, question wording, and the data collection mode.

When the response burden is high, respondents have difficulties answering a questionnaire (Couper and Groves 1996). In establishment surveys, a high response burden is associated with low data quality and high data collection costs (e.g., Bavdaž et al. 2015, Jones 2012, Giesen 2012, Giesen et al. 2011a, Hedlin et al. 2005, Haraldsen and Jones 2007). A high burden can also lead to more data editing and fewer timely responses (e.g., Haraldsen and Jones 2007, Berglund et al. 2013, Giesen 2013) and may reduce respondents' motivation and efforts to answer correctly (Krosnick 1991).

One way to reduce the response burden in establishment surveys may be to change the survey mode from paper to the web. The web mode offers many advantages that can reduce the response burden. However, it can also introduce response burden if respondents are not comfortable with website navigation and forms. In practice, many surveys offer a choice of web or another mode, often paper. The choice of mode may allow respondents to choose their preferred mode, leading to a lower burden; or it may present respondents with another decision they must make, leading to a higher burden, as in Medway and Fulton (2012). Studies of the change in the response burden when moving establishment surveys from paper to the web have found that the introduction of the web mode reduces the response burden (Giesen 2013b, Gravem et al. 2011, Giesen et al. 2009) or has no effect (Snijkers et al. 2007). However, because the questionnaire content and structure in those studies also changed, we cannot draw a definitive conclusion on the effect of web on response burden (Gravem et al. 2011).

To address the shortcomings of previous studies, we conducted an establishment survey with an experimental assignment to the mode: *Paper-only*, *Web-only* or concurrent *Paper and Web mixed mode*. We examine the differences in response burden between modes, and we will answer the following research questions:

Is response burden in an establishment survey lower in the web mode than it is in the paper mode?

Do respondents experience a lower burden if they can choose between the paper and the web mode?

To answer our research questions, we first define what type of response burden we evaluate. Second, by listing the benefits of the web mode, we explain why data collection agencies are interested in using the web mode for their establishment surveys. Third, we provide a literature overview on how response burden is measured. Fourth, we describe the possible effects of paper and the web on response burden, leading us to our hypotheses. Fifth, we describe our data, including our study and experimental design as well as key features of our web survey. Sixth, we describe the models we use to evaluate response burden differences. Seventh, we present our results. Finally, we summarize our results and the limitations of their scope.

2.3 Background

Establishment surveys can impose burden in three ways (Löfgren 2011, Haraldsen et al. 2013). First, each time an establishment is selected for a survey, the establishment is burdened with a response request, and large establishments are selected more often than medium and small establishments (Jones 2012). Second, for those establishments that choose to participate, the participation costs are presumably greater than the benefits to the establishment (Verkruyssen and Moens 2011, Giesen 2011). As a result, establishments may have a low motivation to respond. Third, instrument design introduces burden through questionnaire content and length, the data collection mode (e.g., face-to-face, telephone, paper, web, etc.), the wording of questions and other factors. This paper focuses on burden introduced through instrument design. We refer to this type of burden as *response burden*. Specifically, we focus on the mode as part of the instrument design and

compare the difference in response burden between paper and web modes in establishment surveys.

In the remainder of this section, we provide a short overview of the benefits of web surveys compared to paper surveys. We explain how response burden is conceptualized and measured and the possible effects of paper and web surveys on response burden. We then develop hypotheses regarding how response burden differs between paper and web surveys.

2.3.1 Benefits of web surveys

Although paper and web are both cost-efficient self-administered modes, web offers several advantages over paper. Web surveys reduce or eliminate mailing costs. Many establishment survey invitations can be sent via email; when mail invitations are used, only an invitation letter is sent rather than a large paper questionnaire and return envelope. Furthermore, web surveys reduce data entry costs. These savings usually more than offset potential increases in programming needed to set up the web survey.

The web mode can also increase data quality. Web questionnaires can provide feedback to respondents (Couper 2008, Conrad et al. 2007). If respondents submit an unlikely answer, plausibility checks can ask respondents to re-evaluate their answers, which could reduce the need for data editing. Furthermore, researchers can offer definitions and additional information on how to answer the question. Future web surveys may even include chatbots that can address respondents' questions during the response process (Lagerstøm 2018). Additionally, the web mode can manage calculation and counting tasks, which simplify responses (Giesen et al. 2009, Giesen 2007). Especially in establishment surveys, which often require responses from multiple respondents, web surveys may simplify the response process within the establishment as respondents can distribute a link for a web survey easily via email while a paper questionnaire is more cumbersome to

distribute to multiple respondents. Finally, web surveys enable a complex filter and skip pattern design while only showing items applicable to each respondent.

2.3.2 Conceptualizing and measuring response burden

Bavdaž et al. (2015) summarize three reasons why National Statistical Institutes (NSIs) should consider response burden when designing data collection programs. The first is political: responding to a survey takes time away from an establishment's core business and may decrease competitiveness. The second is methodological: a high burden may reduce data quality and increase data collection costs. The third is strategic: burden can negatively affect the relationship between NSIs and the business community, reducing the motivation to respond to surveys. Therefore, NSIs should monitor and reduce burden to the fullest possible extent (e.g., see European Commission 2011), and burden management has become a key element for NSIs (e.g., see Giesen et al. 2018).

Response burden is a multifaceted concept influenced by motivation, task difficulty, survey effort and respondent perception (Yan et al. 2020). It is often "loosely defined", and Yan et al. argue for a unified concept for response burden. For our study, we follow the conceptualization of actual and perceived response burden, which we find is the most prominent within the establishment survey literature (e.g., Giesen 2013, Berglund et al. 2013, Hedlin 2005, Giesen et al. 2009, Giesen and Burger 2013, Haraldsen and Jones 2007). The literature suggests several indicators to measure actual response burden. Because respondents need time to read, think, and respond to a question, each item in the survey adds to the overall burden (Bradburn 1978). Therefore, questionnaire length is probably the most basic indicator for response burden (see, e.g., Groves, Cialdini and Couper 1992, Van Loon et al. 2003). In our study, we asked respondents how much time they spent answering the questions (see, e.g., Dale et al. 2007, Giesen et al. 2011a, Giesen

2013b). Additional indicators used by NSIs to track response burden imposed on establishments include the following: calls to the service number, requests for help, response rates, and average time for questionnaire completion (Downey et al. 2007, Snijkers et al. 2007, Sear 2011, Giesen et al. 2011a).

Perceived response burden is a subjective measure of respondents' experiences responding to the survey, e.g., as burdensome and time consuming (see, e.g., Haraldsen et al. 2013). It is not the actual time spent taking a survey but the perception of the time and effort of the survey that affects respondents' survey experience and response quality (e.g., Haraldsen and Jones 2007). Many factors can contribute to perceived response burden: structures within the establishment (who has the information needed to respond), the timing of a survey (during a firm's busy period or while a key informant is on vacation), question design, data collection mode, number of survey invitations, difficulty of the response task, and attitudes towards the data collector (Hedlin et al. 2005, Giesen 2013b).

Perceived response burden is often collected with two items. One item asks for the perception of time on a five-point scale, i.e., if respondents perceive the survey as quick or time-consuming. The other item asks for the perception of burden on a five-point scale, i.e., if respondents perceive the questionnaire as easy or burdensome to answer (see, e.g., Dale et al. 2007, Giesen et al. 2011a, Giesen 2013b). We use the same perceived response burden indicators for our study.

Actual and perceived response burdens are conceptually different from each other but positively correlated (Giesen 2013a, Berglund et al. 2013). If respondents perceive a questionnaire as difficult, the actual response burden (time spent) is also likely high (Giesen 2013a). Giesen et al. (2011a) found that 34 of 41 NSIs collect data on actual response burden while 12 collect data on perceived response burden. We examine how assigned mode and mode choice affect both actual burden and perceived burden.

2.3.3 Possible effects of paper and web surveys on response burden

The impact of mode on actual and perceived burden is complex. Each page of a paper questionnaire introduces an additional workload, and respondents may perceive multipage questionnaires as burdensome. Even if not all questions apply to the respondent, the number of pages can make the survey seem overwhelming. Skip instructions in paper questionnaires may not be clear to respondents, and they may have a hard time navigating a paper survey. Web surveys, on the other hand, do not show all questions to the respondent but only those that apply. As a result, respondents never see the entire questionnaire and cannot immediately assess its total length. They also do not need to pay attention to filter instructions, which reduces the respondent's cognitive effort.

On the other hand, the web mode could increase response burden. Respondents with lower online skills may experience a greater burden (Gregory and Earp 2007). A poorly designed instrument can be difficult or frustrating to fill out. Furthermore, even well-designed plausibility checks may increase response burden (Hedlin et al. 2005).

Most NSIs do not use web as a standalone mode but in combination with other modes of survey data collection, often a paper mode. Offering the web in addition to paper may reduce the perceived response burden: faced with a choice of mode, respondents should choose the mode they feel most comfortable responding to and the one that is lower burden for them (Erikson 2007). Lyly-Yrjänäinen and van Houten (2011) propose offering multiple modes to reduce the respondent burden in Eurostat establishment surveys. However, offering multiple modes can overwhelm respondents and reduce response rates (Medway and Fulton 2012). Requiring respondents to choose a mode before they can begin the survey may also impose an additional burden on respondents.

2.3.4 Hypotheses

The above discussion leads us to several hypotheses regarding the relationship between the mode and response burden in establishment surveys. In accordance with the findings from earlier research (Gravem et al. 2011, Giesen et al. 2009, Snijkers et al. 2007), we hypothesize that burden will be lower for respondents assigned to the web mode than for those assigned to the paper mode (hypothesis 1).

Therefore, compared to the paper mode, we expect ...

... a shorter time to complete the questionnaire in the web mode (hypothesis 1.1)

... a lower perceived time in the web mode (hypothesis 1.2)

... a lower perceived burden in the web mode (hypothesis 1.3)

Hypothesis 2 relates to mode choice: when respondents can choose their mode, they are likely to experience a lower burden than respondents who respond in the same mode but were not given a choice. We hypothesize that actual and perceived response burden among those who choose the web mode from a mixed-mode condition are lower than burden among those assigned to the web mode (hypothesis 2.1). Therefore, compared to the assigned web condition, we expect...

... a shorter time to complete the questionnaire by web respondents in the mixed-mode condition (hypothesis 2.1.1)

... a lower perceived time by web respondents in the mixed-mode condition (hypothesis 2.1.2)

... a lower perceived burden by web respondents in the mixed-mode condition (hypothesis 2.1.3)

Similarly, we expect lower actual and perceived response burden for respondents in the paper mode from a mixed-mode condition compared to respondents from the assigned paper condition (hypothesis 2.2). That is, compared to the assigned paper condition, we expect...

... a shorter time to complete the questionnaire by paper respondents in the mixed-mode condition (hypothesis 2.2.1)

... a lower perceived time by paper respondents in the mixed-mode condition (hypothesis 2.2.2)

... a lower burden by paper respondents in the mixed-mode condition (hypothesis 2.2.3)

Although respondents likely use their preferred mode when choosing between paper and web, we should still see differences in response burden between those choosing paper and those choosing web. The features of the web mode described earlier should reduce response burden. Therefore, we hypothesize that actual and perceived response burden will be lower for those who respond via the web in the mixed-mode condition than for those who respond via paper in the mixed-mode condition (hypothesis 3). Therefore, compared to those choosing paper, we expect ...

... a shorter time to complete the questionnaire for those choosing web (hypothesis 3.1)

... a lower perceived time for those choosing web (hypothesis 3.2)

... a lower perceived burden for those choosing web (hypothesis 3.3)

2.4 Data

To examine our hypotheses regarding the differences in response burden between modes, we use data from a German establishment survey. The Institute for Employment Research (IAB) designed this survey to evaluate the effect of the mode on the data quality in establishment surveys.

Overall, 16,000 establishments were sampled from German administrative records. Sample selection was stratified by location (East and West Germany), establishment size class (< 10 employees, 10 - 199 employees, and ≥ 200 employees) and industry class following the German Classification of Economic Activities (Destatis 2008). Establishments already selected for IAB surveys in 2015 were removed from the frame before selection to avoid causing any problems for those ongoing data collection efforts. The removed establishments were random selections from the frame and thus should not bias the sample. However, there are some strata where no unselected establishments remained on the frame. This issue particularly affected the largest size class in which there are few establishments. For this reason, the sample used in this study is not fully representative of the population of establishments, but efforts were made to be as complete as possible given the need to avoid overlap with ongoing surveys. Participation in the survey was voluntary, and the overall response rate was 10.2% (AAPOR RR1) with 1,574 establishments responding.

All sampled establishments were randomly assigned to one of the three mode conditions (*Paper-only*, *Web-only*, *Choice*). To ensure we would have enough cases in all three mode groups and within the two modes in the *Choice group*, we assigned one-fourth of the establishments in our sample to *Paper-only*, one-fourth to *Web-only* and two-fourths to *Choice*.

We prepared two versions of the questionnaire with different topics, number of items and question formats. One version focused on the consequences of the introduction of the federal minimum wage in Germany in 2015. We refer to this version as *Minimum Wage*. Another version contains questions about the effect of increasing digitalization on labor markets. We refer to this version as *Digitalization*. We randomly assigned each sampled establishment to one of the two versions. Therefore, both versions are independent surveys with the same experimental mode design. However, our hypotheses should apply to both questionnaire versions. In fact, seeing similar results over both versions should increase the reliability of our results. All mode groups were invited to participate in the study via a mailed letter. For the *Paper-only* group, we mailed establishments a cover letter with information about the study and a paper questionnaire. Depending on the assigned versions, the number of pages and questions differed slightly. The *Minimum Wage* questionnaire contained 74 questions on 20 pages. In contrast, the *Digitalization* questionnaire had 69 questions on 19 pages, printed in a 20-page booklet. Therefore, the difference in page volume between both versions was negligible.

For the *Web-only* group, we sent establishments a cover letter with information about the study, a link and the request to fill out our online questionnaire. To isolate mode effects, we took care to ensure that the paper and web questionnaire were visually similar to each other. However, the web mode offers functionalities that may reduce response burden, as discussed above. We implemented six web survey functionalities. First, the web survey presented questions in a paging design (one question on each page) so that respondents would not miss a question. Second, the web survey used automatic skips, i.e., questions that did not apply to respondents were not shown. Third, the question about the number of different employment groups automatically summed and displayed the total number of employees. Fourth, we implemented plausibility checks. For instance, if the respondent stated that the regular weekly working hours were greater than the legal limit of 48 hours,

the web survey prompted an error message in red that asked respondents to re-evaluate their answer. The number of plausibility checks differed by questionnaire version: *Minimum Wage* contained up to 13 plausibility checks, and *Digitalization* contained up to five plausibility checks. Fifth, at the end of each section, respondents were able to print the questionnaire section with their responses for their own documentation.

Sixth, the web survey contained an index that allowed respondents to navigate to specific sections. The index indicated the structure of the questionnaire and showed the headings for each section (see Figure 2.1). After finishing a section, the web survey redirected respondents to this index page. The index page gave respondents an understanding of what part of the questionnaire should be answered by whom in the establishment. In establishment surveys, respondents sometimes do not have the information required to answer all questions. Therefore, they require help from colleagues to answer some questions.

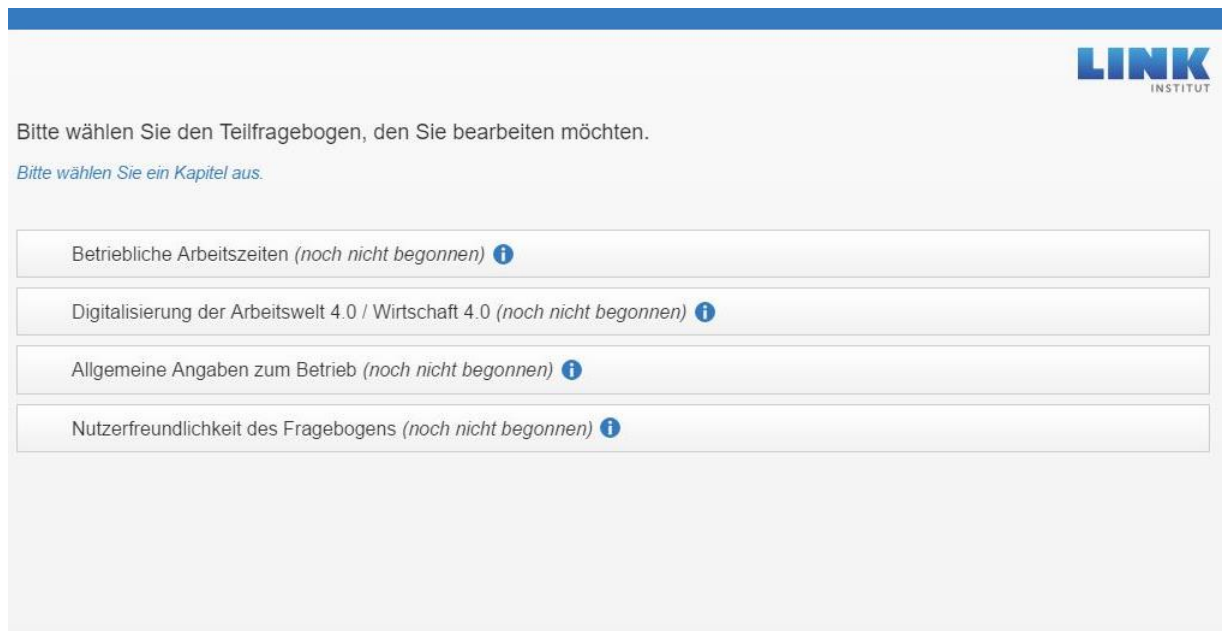
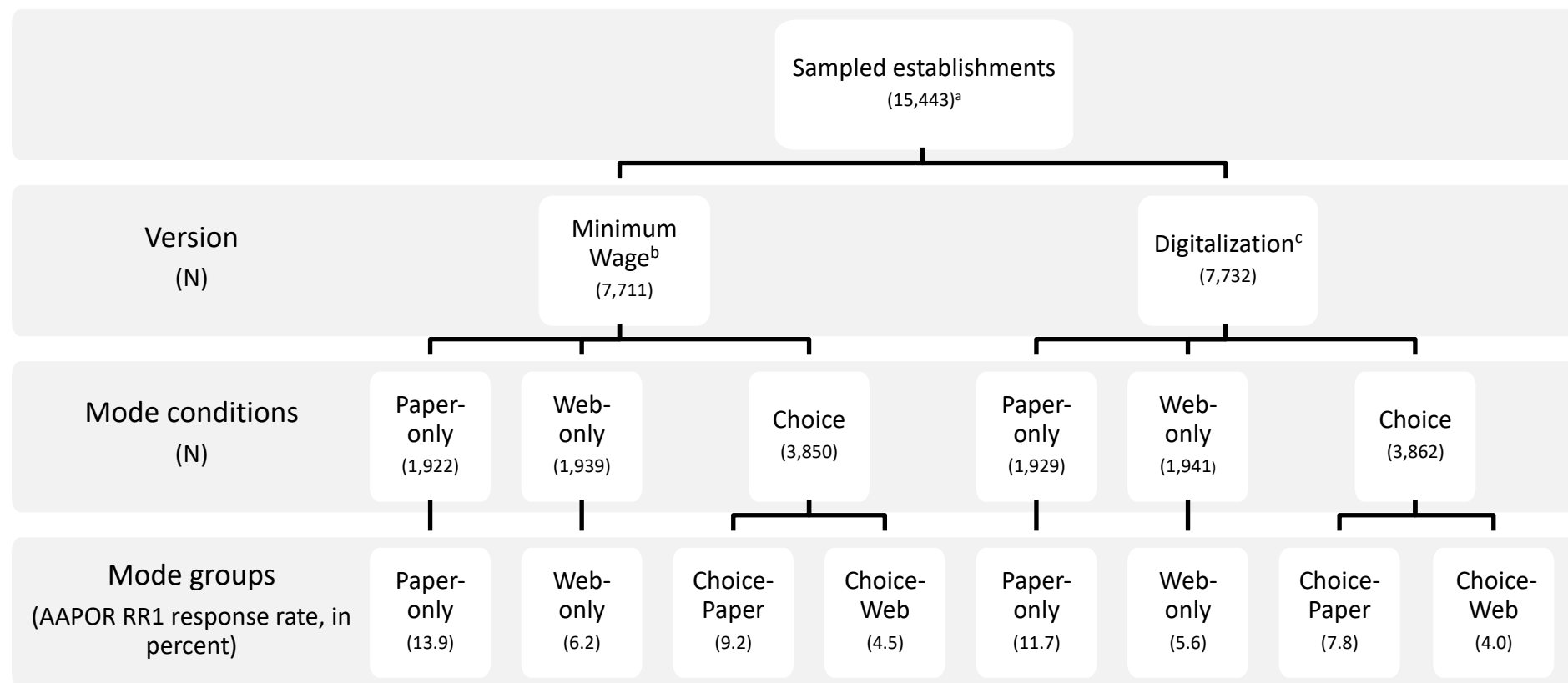


Figure 2.1: Index page for the web survey in the Digitalization version

For the establishments in the *Choice* groups, we sent a cover letter and the same paper questionnaire as in the *Paper-only* group. The cover letter offered a web link and presented the option to choose between the paper and web modes.



^a All Ns exclude 557 cases found to be ineligible.

^b Overall AAPOR RR1 for *Minimum Wage* is 11.9%.

^c Overall AAPOR RR1 for *Digitalization* is 8.5%.

Figure 2.2: Experimental Assignment and Response Rates (RR)

Figure 2.2 shows the response rates (RR) for our three mode groups and two versions. In the figure, *Choice-Paper* refers to cases that chose to respond via the paper mode in the mixed-mode condition. *Choice-Web* refers to those that responded on the web. In both versions, response rates in the *Web-only* and *Choice-Web* groups are smaller than those in the *Paper-only* and *Choice-Paper* groups. The response rates are also lower in all conditions for the *Digitalization* survey than the *Minimum Wage* survey (8.5% vs. 11.9%). Furthermore, we find that compared to *Paper-only*, the *Choice* group is not different in terms of response rates (13.7% vs 13.9% in the *Minimum Wage* survey; 11.8% vs 11.7% in the *Digitalization* survey). (To calculate the response rate for *Choice-Paper* and *Choice-Web*, we split the response rate of *Choice* into the proportion of *Choice-Paper* and *Choice-Web*, i.e., $RR_{Choice} = RR_{Choice-Paper} + RR_{Choice-Web}$.) These results contradict findings from meta-analyses where offering a choice between modes is burdensome enough to not participate (Medway and Fulton 2012). However, the meta-analysis did not include establishment surveys.

To check whether respondents in each mode group differ from each other, a nonresponse analysis for the variables location (East and West Germany), establishment size class (< 10 employees, 10 - 199 employees, and ≥ 200 employees) and industry was conducted (see Haas et al. 2016). No systematic differences in nonresponse patterns between the mode groups were found.

Involving other people and managing the response process can be a burden to respondents. Overall, 16.2% of our respondents reported that they had help answering the questionnaire. Concerning the proportion of multiple respondents, a chi-squared test suggests no differences between the mode groups and questionnaire versions ($\chi^2_{7, N=1,663} = 5.6, p < 0.585$).

2.5 Methods

To evaluate the differences in response burden between survey modes, we use median and ordered regression models. The dependent variables in all models are measures of response burden. The independent variables are the experimental conditions (mode and topic) and control variables about the establishments (size class, industry, and East versus West Germany).

2.5.1 Response burden variables

We measure actual and perceived burden with three questions (Dale et al. 2007) asked at the end of the questionnaire. First, we asked respondents to estimate the *time they needed to complete* the questionnaire. The question required an answer in hours and minutes and has been used as a measure of actual burden in earlier studies (e.g., Dale et al. 2007, Giesen 2013, Berglund et al. 2013). However, as respondents retrospectively estimate time and do not actively measure it, our measure of actual burden is not as objective as the literature may suggest. Second, we asked respondents to rate the *perceived time* taken on a 5-point scale from “very quick” to “very time consuming”. Third, we asked respondents to rate the *burden of the survey* on a 5-point scale ranging from “very easy” to “very burdensome”. For the sake of simplicity, we will refer to these two variables as *perceived burden indicators*. Furthermore, we recode our scales from 5 points to 3 points (0, 1 and 2) by collapsing the two categories at each end. The results are not substantially different between the five- and three-point scales, but the 3-point scale makes it easier for the reader to interpret the results. For the full wording of the three burden questions and response options, see Appendix Table 2.4.

2.5.2 Independent variables

We have four mode groups, i.e., *Paper-only*, *Web-only*, *Choice-Paper*, and *Choice-Web*, which are our independent variables of interest. We can test our three hypotheses by comparing the four groups. First, we compare *Paper-only* and *Web-only* to test whether response burden is lower for web in an establishment survey (hypothesis 1). Second, we compare *Paper-only* and *Choice-Paper* and also *Web-only* and *Choice-Web* to test whether having the chance to choose a mode affects response burden (hypothesis 2). Third, we compare *Choice-Paper* and *Choice-Web* to test whether response burden is lower among respondents who opted for the web mode (hypothesis 3).

The models also control for the number of questions the respondent answered. Due to filters and skip patterns, the number of questions each respondent answered was not tightly controlled, even within the same questionnaire version. Therefore, we introduce the variable *number of applicable items* for each respondent. This variable counts the number of items respondents should have answered from the start of the interview until the response burden questions. In the last section, we asked respondents which questionnaire sections they answered themselves (as opposed to which ones a colleague answered). If they reported that more than one person answered the questionnaire, we consider only the number of items that the final respondent answered in the model because that respondent was the one who answered the burden questions. Furthermore, we include the indicator of more than one respondent in the model as a dummy variable.

The models also control for location, size and industry to account for possible selection bias between modes and to increase the precision of our estimates.

2.5.3 Models

To evaluate our hypotheses on response burden differences between modes, we use multivariate regression models. We ran a model for each of our three response burden variables: *time to complete the questionnaire*, *perceived time* and *perceived burden*. Furthermore, we ran our models for each questionnaire version separately. Therefore, we have six models. Because we do not claim to represent the population of establishments, all analyses are unweighted. Each model does include the three stratification variables as controls in all models; however, they are the only variables that influence the weights. Controlling for the components of sample weights is an alternative to the use of weights in regression analyses (Gelman 2007).

Because the dependent variables have different scales, we use different models. Our response burden variable *time to complete the questionnaire* has large outliers (see **Fehler! Verweisquelle konnte nicht gefunden werden.**). For this reason, we use a median regression that is less susceptible to being influenced by very short and very long times than an ordinary least squares regression (e.g., Cameron and Trivedi 2005):

$$y_i = M_i' \beta_M + X_i' \beta_X + \varepsilon_i \quad (1)$$

where y_i is the time to complete the questionnaire for a questionnaire version, M_i' is the mode group, X_i' are the controls and ε_i are the unobserved variables or errors.

Using a median regression, we assume that $MED(\varepsilon_i | M_i', X_i') = 0$, which implies that:

$$MED(y_i | M_i', X_i') = M_i' \beta_M + X_i' \beta_X \quad (2)$$

Table 2.1: Summary statistics for the time to complete the questionnaire in minutes by questionnaire version and mode group

	Minimum Wage					Digitalization				
	N	Mean	Me- dian	Min	Max	N	Mean	Me- dian	Min	Max
Paper-only	272	34.4	30	5	180	190	48.9	35	10	240
Web-only	116	26.2	20	2	120	91	38.2	30	5	165
Choice-Pa- per	355	37.5	30	5	210	245	55.4	30	5	1,440
Choice-Web	171	32.1	20	1	210	123	44.9	30	1	1,200
Overall	914	34.2	30	1	210	649	49.1	30	1	1,440

The two *perceived burden* variables are ordinal scales, and we use ordinal logistic regression models with these variables (e.g., see Cameron and Trivedi 2005, p. 519 f.) and adapt our model as follows:

$$y_i^* = M_i' \beta_M + X_i' \beta_X + \varepsilon_i \quad (3)$$

$$Y_i = \begin{cases} 0 & \text{if } y_i^* \leq \alpha_0 \\ 1 & \text{if } \alpha_0 < y_i^* \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < y_i^* \end{cases} \quad (4)$$

where y_i^* is one of our perceived burden indicators and α_i the threshold parameters that are obtained by maximizing the log-likelihood. We calculate the marginal effects in the probabilities as follows:

$$\frac{\delta \Pr [y_i = j]}{\delta M_i'} = \{F'(\alpha_{j-1} - (M_i' \beta_M + X_i' \beta_X)) - F'(\alpha_j - (M_i' \beta_M + X_i' \beta_X))\} \beta_M \quad (5)$$

where F' denotes the derivative of the cumulative distribution function of ε_i .

The independent variables in all models are the same. Table 2.2 summarizes the six models.

Because we focus on the differences between modes, we report only the linear prediction of the median time from the median regression ($\frac{\text{MED}(y_i | M_i', X_i')}{\delta M_i'} = \beta_M$) and the predicted probabilities from the ordinal logistic regression (*Equation (5)*) for our mode groups.

Table 2.2: Summary of the six models for evaluating the response burden

Model	Dependent Variable	Questionnaire Version	Model Type	Independent Variables
1	Time to complete	Minimum Wage	Median regression	• Mode • Number of applicable items • Establishment size • Industry • Region • Multiple respondents (Yes/No) • Interaction of mode with each of the above (except mode)
2	Time to complete	Digitalization		
3	Perceived time	Minimum Wage		
4	Perceived time	Digitalization	Ordinal logistic regression	
5	Perceived burden	Minimum Wage		
6	Perceived burden	Digitalization		

The results of each model provide information supporting or rejecting our hypotheses. Running the models on the two questionnaire versions separately provides us with information about whether our results hold across both survey topics. Support for hypotheses 1.1 to 1.3 (response burden is lower in the web mode than in the paper mode) will be seen

by comparing the coefficients of the mode indicators for Paper-only and Web-only. For the time to complete the questionnaire, we expect to see a lower estimated time for the Web-only group. For both perceived indicators, we expect to see higher predicted probabilities for the categories “quick” (perceived time) and “easy” (perceived burden) in the Web-only group. For hypotheses 2.1.1 to 2.1.3, we compare the coefficients of Choice-Web against Web-only; and for hypotheses 2.2.1 to 2.2., we compare Choice-Paper and Paper-only. We expect a lower burden in the Choice conditions than in the Only conditions. For hypotheses 3.1 to 3.3, we compare the coefficients of Choice-Web against Choice-Paper. We expect all three models to indicate lower burden in Choice-Web than Choice-Paper.

2.6 Results

Before presenting the results of our hypothesis tests, we examine the burden within the two questionnaire versions with the three response burden indicators (*time to complete the questionnaire*, *perceived time* and *perceived burden*).

On average, respondents to the *Minimum Wage* version needed less time to *complete the questionnaire* (34 vs. 49 minutes) than respondents in the *Digitalization* version. As the data for the *time to complete the questionnaire* is not normally distributed (see Table 2.1), we cannot conduct a two-sample t-test. However, a nonparametric equality-of-medians test (see Snedecor and Cochran 1989) shows that *complete time* ($\chi^2_1, N=1,563 = 50.1, p < 0.001$) is different between the two versions.

Table 2.3 shows the descriptive results of our perceived time indicators for each questionnaire version independent of the mode. We use a chi-squared test to examine differences in the perceived time indicators between our questionnaire versions. Overall, the *Digitalization* version is perceived as more time consuming ($\chi^2_2, N=1,668 = 32.0, p < 0.001$) and burdensome ($\chi^2_2, N=1,660 = 54.6, p < 0.001$) than the *Minimum Wage* version (see Table

2.3). Because burden is very different in the two questionnaire versions, we run separate models for the two versions in the rest of the paper.

Table 2.3: Proportions of perceived time and burden by questionnaire version.

Perceived time*	Minimum Wage (N= 967)	Digitalization (N= 701)
Quick	57.3	44.4
Neither	34.1	40.5
Time consuming	8.6	15.0

Perceived burden**	Minimum Wage (N= 962)	Digitalization (N= 698)
Easy	66.6	49.0
Neither	29.0	42.1
Burdensome	4.4	8.9

* $\chi^2 = 32.0$, $p < 0.001$; ** $\chi^2 = 54.6$, $p < 0.001$

Hypothesis 1: The response burden in the Web-only mode is lower than that in the Paper-only mode.

We hypothesized that the web mode leads to a lower response burden. We test this hypothesis using the six models described in the methods section. For the time to complete the questionnaire, we expect to see a lower estimated time for the *Web-only* group than for the *Paper-only* group. For both *perceived indicators*, we expect to see higher predicted probabilities for the categories “quick” (perceived time) and “easy” (perceived burden) in the *Web-only* group compared to the *Paper-only* group.

Figure 2.3 compares the marginal effects of the four mode conditions on the median *time to complete the questionnaire* for the *Minimum Wage* version. At the median, respondents

assigned to the *Web-only* group needed 5.5 fewer minutes to complete the questionnaire than respondents in the *Paper-only* group (based on self-reported completion time; $F_{1, 886} = 4.9$, $p < 0.013$). As the *time to complete the questionnaire* is lower in the web group, the results support hypothesis 1.1 that the web mode has a lower actual burden than the paper mode.

Figure 2.4 shows the average predicted probabilities from the ordinal logistic regression model for respondents' perceived time and burden over the four mode groups in the *Minimum Wage* version. The predicted probabilities provide us with a measure of how the respondents perceived responding to the survey mode while controlling for our independent variables (size, industry, number of applicable items and region). The left panel shows the results for *perceived time*, and the right panel shows the results for *perceived burden*. Our model predicts similar probabilities for *Paper-* and *Web-only* for perceived time ($\hat{p}_{\text{quick}}=0.54-0.61$, $\hat{p}_{\text{neither}}=0.32-0.37$, and $\hat{p}_{\text{time consuming}}=0.07-0.09$) and perceived burden ($\hat{p}_{\text{easy}}=0.64-0.72$, $\hat{p}_{\text{neither}}=0.25-0.31$, and $\hat{p}_{\text{burdensome}}=0.03-0.05$). We see no variation between *Web-only* and *Paper-only* for either of our *perceived burden indicators* ($\chi^2_{4, N=409} = 3.2$, $p=0.53$ for *perceived time* and $\chi^2_{4, N=409} = 0.7$, $p = 0.95$ for *perceived burden*). The results from the two *perceived burden indicators* do not support hypotheses 1.2 and 1.3.

The results for the *Digitalization* version are similar (see Figure 2.5 for the *time to complete the questionnaire* and see Figure 2.6 for *perceived time indicators*). Figure 2.5 shows that the estimated time for completing the questionnaire is 10 minutes lower in *Web-only* ($F_{1, 615} = 6.0$, $p = 0.007$) and supports hypothesis 1.1 that response burden is lower for the web mode. In Figure 2.6, there is no variation in the marginal predicted probabilities between *Web-only* and *Paper-only* for *perceived time* ($\chi^2_{4, N=298} = 0.1$, $p = 1.0$) and *perceived burden* ($\chi^2_{4, N=295} = 0.5$, $p = 0.97$). Therefore, our results do not support hypotheses 1.2 and 1.3 that the perceived response burden is lower in the web mode.

Hypothesis 2: The burden is lower when respondents choose a mode than when that mode is assigned.

We hypothesized that the possibility of choosing one's preferred mode lowers the burden of respondents. Therefore, the estimated time should be smaller for the (2.1) *Choice-Paper* group than for the *Paper-only* and for the (2.2) *Choice-Web* group than the *Web-only*.

For the *Minimum Wage* version (see Figure 2.3), the difference between *Paper-only* and *Choice-Paper* in the *time to complete the questionnaire* is 0.4 minutes. In the group comparison of *Web-only* and *Choice-Web*, we find a -0.6-minute difference. For the *Digitalization* version (see Figure 2.5), there is no difference in the *time to complete the questionnaire* between the *Paper-only* and *Choice-Paper* groups and between the *Web-only* and *Choice-Web* groups.

For our *perceived burden indicators* (see Figure 2.4 for the *Minimum Wage* version and Figure 2.6 for the *Digitalization* version), there is no variation in the predicted probabilities between our *Choice* and *Only* groups (see Appendix Table 2.5 for the joint χ^2 values). Overall, we find no support for our hypotheses 2.1.1 to 2.2.3.

Hypothesis 3: The response burden in the web mode is lower than that in the paper mode (mode choice).

Our third hypothesis is similar to the first, but it compares paper and the web when a choice is offered. We hypothesized that among those respondents given a choice, the web mode should have a lower response burden than the paper mode.

In the *Minimum Wage* questionnaire, *Choice-Web* respondents needed 6.4 fewer minutes to complete the questionnaire ($F_{1, 886} = 9.5, p = 0.001$) (see Figure 2.3) than *Choice-Paper*

respondents. In the *Digitalization questionnaire*, the differences were larger: the estimated median *time for completing the questionnaire* is 10 minutes lower in *Choice-Web* ($F_{1, 625} = 7.6$, $p = 0.003$) (see Figure 2.5).

Examining Figure 2.4 and Figure 2.6, we see no variation in the predicted probabilities between *Choice-Paper* and *Choice-Web* (see Appendix Table 2.5 for the joint χ^2 values). Therefore, we find mixed support for our hypothesis: when offered a choice, those choosing the web mode have a lower estimated time for completing the questionnaire (hypothesis 3.1) but there is no difference in the perceived burden (hypotheses 3.2 and 3.3).

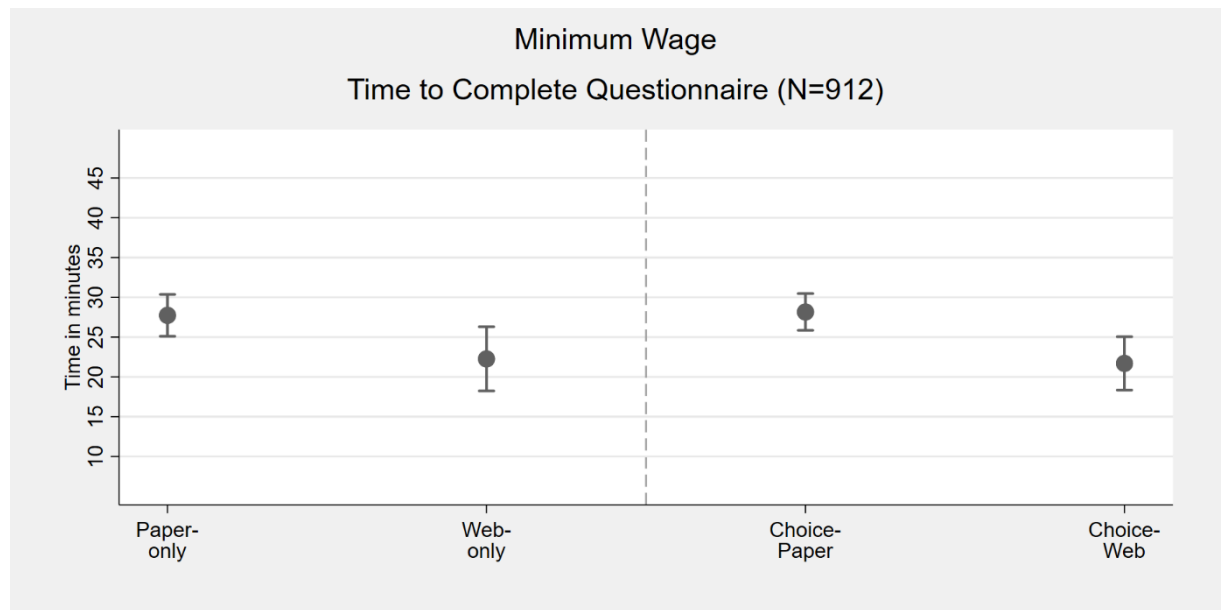


Figure 2.3: Linear prediction of the estimated median time to complete the questionnaire in minutes for the Minimum Wage questionnaire (bars show 95% confidence intervals)

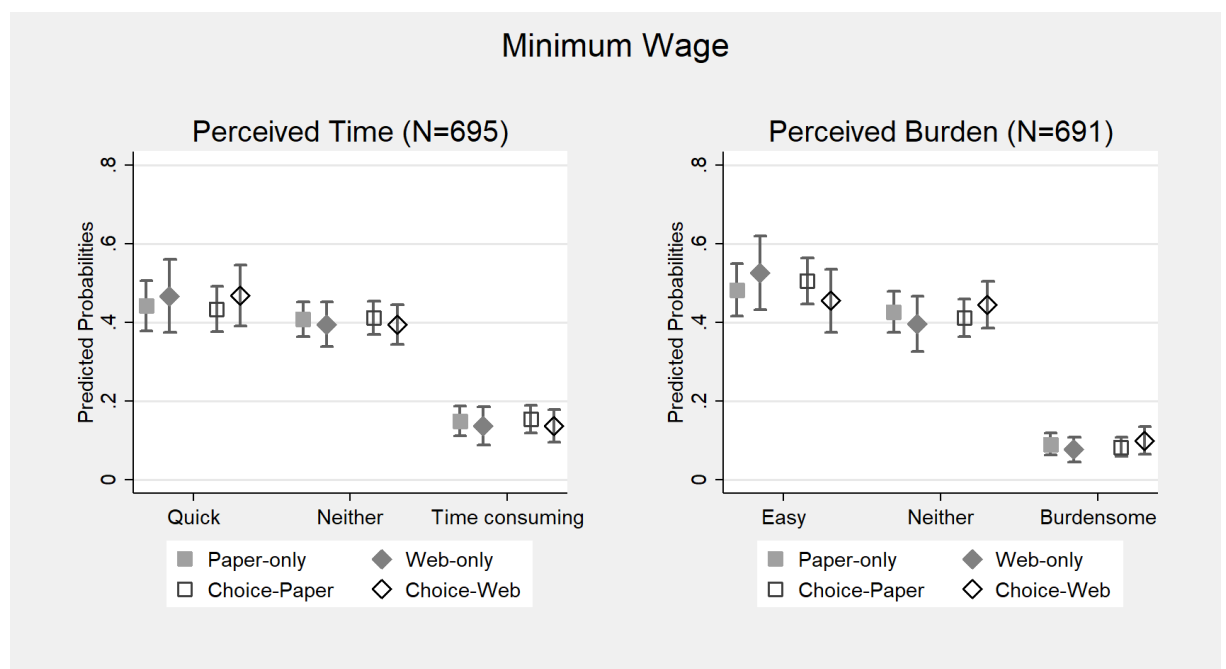


Figure 2.4: Predicted probabilities from the ordinal logistic regression model for perceived time and perceived burden in the Minimum Wage version by mode group (bars show 95% confidence intervals)

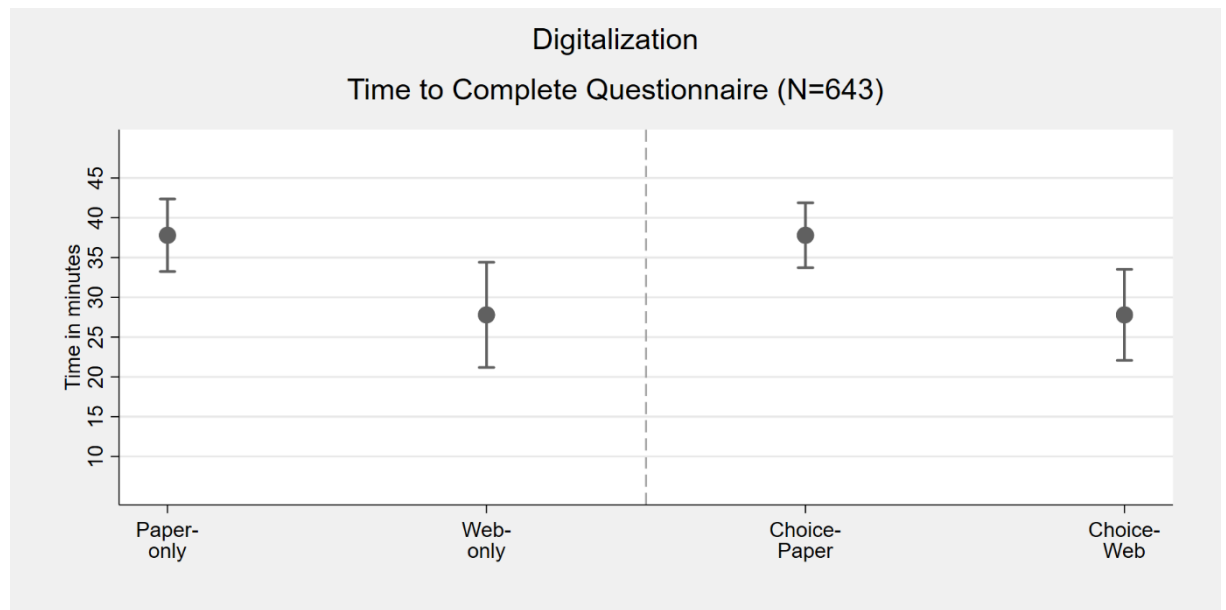


Figure 2.5: Linear prediction of the estimated median time in minutes to complete the questionnaire for Digitalization questionnaire (bars show 95% confidence intervals)

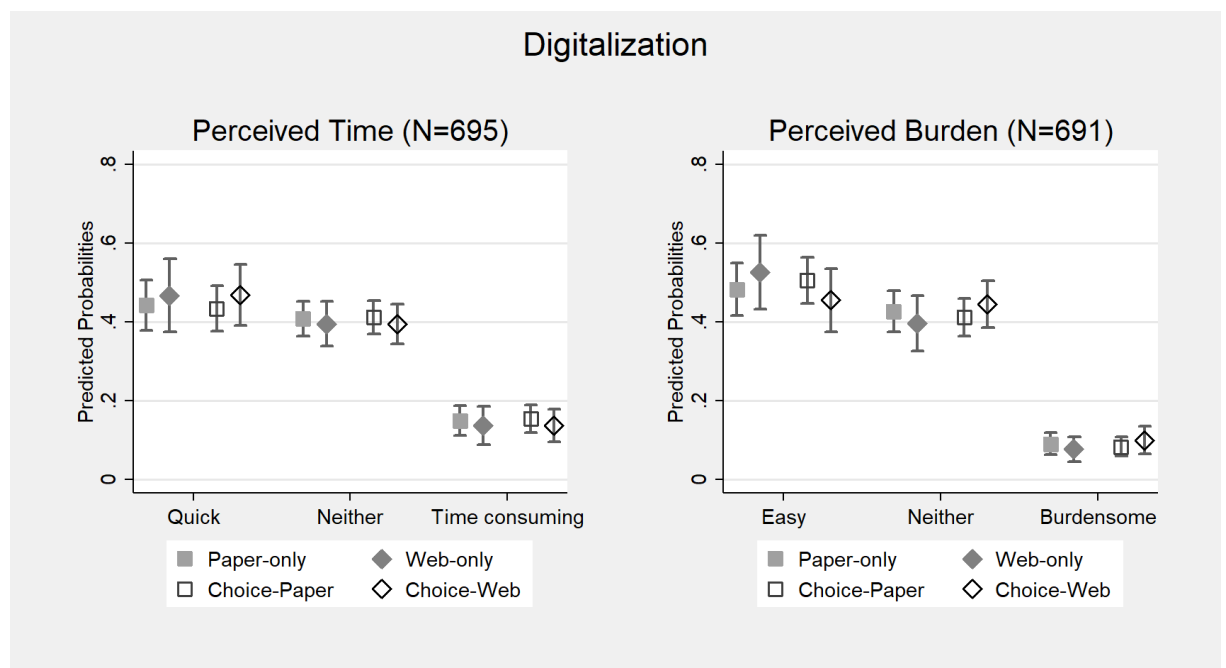


Figure 2.6: Predicted probabilities from the ordinal logistic regression model for perceived time and perceived burden in the Digitalization version by mode group (bars show 95% confidence intervals)

2.7 Conclusion

We designed this study to determine the differences in response burden between paper and web modes in a German establishment survey. We designed two surveys with the same experimental mode groups. To evaluate response burden, we used three measures of burden (estimated time to complete the questionnaire, perceived time and burden) and four mode comparisons (*Paper-only* vs. *Web-only*, *Choice-Paper* vs. *Paper-only*, *Choice-Web* vs. *Web-only*, *Choice-Paper* vs. *Choice-Web*) to answer our research questions about whether response burden is lower in an establishment web survey and whether respondents feel less burdened if they can choose between paper and web modes.

This study has shown that web respondents, whether they were offered the web as a standalone mode or concurrently with a paper questionnaire, have a lower median time to complete the questionnaire compared to a paper questionnaire. These results held when respondents chose the web mode and when they were assigned to the web mode.

We found no evidence of a difference in either measure of perceived burden between the paper and web modes. As we have mentioned at the beginning of this paper, response burden is a multifaceted concept. It is important to note that perceptions of burden could be affected by factors other than time. For instance, a questionnaire that seems relevant and straightforward to respondents might be less burdensome than a shorter but more difficult instrument.

Our results suggest that offering respondents the choice of their preferred mode has no effect on response burden compared to a single-mode setting. Therefore, concerning response burden, the web mode is a cost-effective alternative to the paper mode¹. Furthermore, the results of our study are consistent across two different topics: *Minimum Wage*

¹ Nevertheless, the reader should be aware that the web mode may have a considerably lower response rate than a paper questionnaire.

and *Digitalization*. Therefore, our results may also be applicable to other surveys.

A reason why we find a lower estimated time for the time to complete the questionnaire may be that the web mode has an automatic questionnaire flow and does not show unnecessary questions to the respondents. However, the estimated time to complete could also indicate that web respondents are more satisfied than paper respondents. Future research needs to address this question.

One could argue that the lower response rate in the web survey is a sign that respondents find that mode more burdensome. However, there are several possible explanations for the lower response rate in the web mode. First, while the paper group received an invitation letter and a 20-page questionnaire, the web group received only a one-page invitation letter, which is easier to overlook. Second, the paper questionnaire may have served as a visible reminder to complete the survey in a way that the one-page letter did not. Third, we have anecdotal evidence from our pretest that some respondents had trouble entering the survey link in their web browser. Therefore, contact persons in the web group may have failed to participate because they could not access the web survey, a challenge that the contact persons in the paper group did not have to overcome. Response burden may not be the driving issue for lower response rates. However, researchers planning to use the web mode for their establishment survey should remember that administrating a web survey comes at a cost of lower response rates.

Finally, we need to consider a number of important limitations:

First, the generalizability of our results is limited to surveys with similar lengths and formats. Our results are especially limited to our web survey design. Web surveys with a different design may perform differently as they may have functions or design characteristics that impact the response burden. We designed our web survey to be visually very

similar to the paper questionnaire. However, important design features may reduce respondents' burden in a web survey. Further research might explore how to reduce the response burden in the web mode.

Second, participation in the surveys used in this study was not mandatory. However, for a large proportion of establishment surveys, participation is required by law. Voluntary establishment surveys are likely to exclude establishments that are not motivated to respond or that anticipate a high response burden. Unfortunately, we can only speculate about the relationship between the anticipated response burden, the response rate and our mode groups. Inviting establishments to participate a web survey may exclude respondents who are not very savvy in using digital technologies and therefore decide not to participate. Establishments in the paper group may have cross-read the questionnaire or even started to respond but decided to not respond. As the choice between several modes is likely to overwhelm respondents to not respond at all (Medway and Fulton 2012), the choice of mode in a mandatory survey may add a perceived burden to respond. In all three mode groups, we may have found a higher response burden if the establishment survey would have been mandatory.

Third, there may be establishments with no internet access or with internal security guidelines that block web surveys or render them poorly (Harrel, Yu and Rosen 2007). Furthermore, establishments may have problems logging in, finding the website or navigating the survey (Bremner 2011, Gregory and Earp 2007). Therefore, our web respondent sample may be biased by an unknown coverage error.

Fourth, offering a paper, web or paper/web survey may recruit different kinds of respondents. Therefore, our respondent sample may be biased by mode-introduced nonresponse not visible in the data. However, the fact that the differences between paper and the web in the *Only groups* and paper and the web in the *Choice groups* are similar and the fact

that our findings are consistent over two questionnaire versions makes us somewhat confident in the validity and reliability of our results.

Fifth, our results only consider German establishments. The results may change in establishment populations with higher or lower digitalization rates or with higher or lower internet penetration rates. Furthermore, we can link our results only to establishments that finished the survey but not to all invited establishments.

Sixth, we only consider the effect of the paper and web modes and not any other mode; instrument design; or interaction between instrument designs, respondent characteristics and establishment structures such as size. Especially in relation to the mode, instrument design decisions, respondents' characteristics and establishment structures can interact with each other. As we know from surveys of individuals, younger, more affluent, and higher educated respondents prefer the web mode over the paper mode (Kaplowitz et al. 2004, Kwak and Radler 2002, Messer and Dillman 2010, Millar et al. 2009). Similar effects may occur in establishment surveys. Our sample does not allow testing for these interactions as the number of cases is insufficient. The interaction of the web mode with other survey properties, respondent characteristics and establishment structures should be evaluated in future research.

Seventh, independent of the mode, first-time respondents must become familiar with the survey instrument. Against this background, respondents will develop individual best practices on how to interact with the survey instrument, i.e., they improve when responding to a mode each time they participate. Therefore, we may see a change in the response burden over time. Future research should assess whether panel participation affects the response burden and whether the response burden decreases or increases over time in the web mode.

Eighth, we used a postal letter as the mode of contact to invite establishments in each mode group to participate. Using a different means for contact, e.g., email, may affect respondents' perceived burden. To access a web survey, respondents usually use a link. If the link is provided within an email, respondents only need to click on that link to access the web survey. If the link is provided on a paper invitation letter, respondents should type the link into their browser search bar to access the web survey, which takes more effort than just clicking on a link. Therefore, in terms of the response burden, contacting establishments with postal letters may increase the burden.

Although these limitations seem numerous, our results provide important insights into the effect of the web mode on the response burden in establishment surveys. Moreover, we are convinced that the validity of our findings is very high due to our rigorous experimental manipulations. In addition, our findings are consistent across two different surveys, which increases the reliability of our results. Our study provides important findings for the development and design of establishment surveys in the online era. Even if the perceived response burden (for respondents) is not lower in the web mode, web surveys are cost effective and enable features that help to improve data quality. Our findings about response burden, combined with the lack of difference in the response rates between the Paper-only and the Choice conditions, lead us to recommend that surveys should offer establishments a choice of paper and web modes.

References

- AAPOR (2016). Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys. Available at: http://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf. (accessed March 2019).
- Bavdaž, M., Giesen, D., Černe, S. K., Löfgren, T. and Raymond-Blaess, V. (2015). Response burden in official business surveys: Measurement and reduction practices of national statistical institutes. *Journal of official statistics*, 31: 559-588. DOI: <https://doi.org/10.1515/jos-2015-0035>
- Berglund, F., Haraldsen, G. and Kleven, Ø. (2013). Causes and consequences of actual and perceived response burden based on Norwegian data. In *Comparative report on integration of case study results related to reduction of response burden and motivation of business*, edited by D. Giesen, M. Bavdaž, and I. Bolko, 29-35.
- Bradburn, N. M. (1978). Respondent burden. In *Proceedings of the Section on Survey Research Methods Section: American Statistical Association*, August 14-17, 1978. 35-40. San Diego, CA: American Statistical Association. Available at: <http://www.asasrms.org/Proceedings/y1978f.html> (accessed March 2021).
- Bremner, C. (2011). An investigation into the use of mixed mode data collection methods for UK business surveys. In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž, March 22-23, 2011. 217-220. Heerlen, The Netherlands: Statistics Netherlands. Available at: <http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACEAD995C/0/2011proceedingsblueets.pdf> (accessed March 2019).
- Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

- Conrad, F. G., Schober, M. F. and Coiner, T. (2007). Bringing Features of Human Dialogue to Web Surveys. *Applied Cognitive Psychology*, 21: 165–187. DOI: <https://doi.org/10.1002/acp.1335>
- Couper, M. P. (2008). *Designing Effective Web Surveys*. Cambridge: Cambridge University Press.
- Couper, M. P. and Groves, R. M. (1996). Household-level determinants of survey nonresponse. *New Directions for Evaluation*, 70: 63-79. DOI: <https://doi.org/10.1002/ev.1035>
- Dale, T., Erikson, J., Fosen, J., Haraldsen, G., Jones, J. and Kleven, O. (2007). *Handbook for monitoring and evaluating business survey response burdens*. Luxembourg: Eurostat.
- Destatis. (2008). *Klassifikation der Wirtschaftszweige*. Wiesbaden, Germany: Statistisches Bundesamt. Available at: <https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/klassifikation-wz-2008.html> (accessed June 2019)
- Downey, K., McCarthy, D. and McCarthy, W. (2007). Encouraging the Use of Alternative Modes of Electronic Data Collection: Results of Two Field Studies. In *Proceedings of the Third International Conference on Establishment Surveys*, June 18-21, 2007. 517-524. Montréal, Canada. Available at: <https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000168.PDF> (accessed September 2019)
- Erikson, J. (2007). Effects of offering web questionnaires as an option in enterprise surveys. In *Proceedings of the Third International Conference on Establishment Surveys*, June 18-21, 2007. 1431-1435. Montréal, Canada. Available at: <https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000168.PDF> (accessed September 2019)

- European Commission. (2011). *European Statistics Code of Practice for the National and Community Statistical Authorities*. Adopted by the European Statistical System Committee, September 28, 2011. Available at: <https://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15> (accessed September 2019)
- Gelman, A. (2007). "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22(2): 153–164. DOI: <https://doi.org/10.1214/088342306000000691>
- Giesen, D. (2013a). Causes and Consequences of Actual and Perceived Response Burden Based on Dutch Data. In *Comparative report on integration of case study results related to reduction of response burden and motivation of business*, edited by D. Giesen, M. Bavdaž, and I. Bolko, 33-39.
- Giesen, D. (2013b). Reducing Response Burden by Questionnaire Redesign. In *Comparative report on integration of case study results related to reduction of response burden and motivation of business*, edited by D. Giesen, M. Bavdaž, and I. Bolko, 63-68.
- Giesen, D. (2012). Exploring Causes and Effects of Perceived Response Burden. In *Proceedings of the Fourth International Conference on Establishment Surveys*, June 11-14, 2012. Montréal, Canada. Available at: <https://ww2.amstat.org/meetings/ices/2012/papers/302171.pdf> (accessed September 2019)
- Giesen, D. (2011). Burden reduction by communication. In *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*, edited by D. Giesen, 33-42.
- Giesen, D. (2007). Does mode matter? Comparing response burden and data quality of paper and an electronic business questionnaire. In *Proceedings of the 6th Conference on Questionnaire Evaluation Standards (QUEST)*, April 24-26, 2007. 150-161. Ottawa,

Canada. Available at:

<https://wwwn.cdc.gov/QBank/QUEST/2007/QUEST%202007%20Proceedings-all%20papers.pdf> (accessed September 2019)

Giesen, D., Vella, M. and Brady, C. (2018). Response Burden Management for Establishment Surveys at Four National Statistical Institutes. *Journal of Official Statistics*, 34(2): 397-418. DOI: <https://doi.org/10.2478/jos-2018-0018>

Giesen, D. and Burger, J. (2013). Measuring and understanding response quality in the Structural Business Survey questionnaires. In *Proceedings of the European Establishment Statistics Workshop*, September 9-11, 2013. Nuremberg, Germany. Available at: <http://doku.iab.de/fdz/events/2013/Session5%20Giesen.pdf> (accessed September 2019)

Giesen, D., Bavdaž, M. and Haraldsen, G. (2011). Response burden measurement: Current diversity and proposal for moving towards standardisation. In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž, March 22-23, 2011. 125-134. Heerlen, The Netherlands: Statistics Netherlands. Available at: <https://www.cbs.nl/-/media/imported/documents/2011/14/2011-4-4-4-2-giesen-et-al-presentation-blue-ets-2011.pdf> (accessed March 2021).

Giesen, D., Morren, M. and Snijkers, G. (2009). The effect of survey redesign on response burden: An evaluation of the redesign of the SBS questionnaires. *Paper presented at the European Survey Research Association Conference 2009*, Warsaw, June 29-July 3, 2009. Warsaw, Poland. European Survey Research Association.

Groves, R. M., Cialdini, R. B. and Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56: 475-495.
<https://doi.org/10.1086/269338>

- Gregory, G. and Earp, M. (2007). Evolution of Web at USDA' National Agricultural Statistics Service. In *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, June 18-21, 2007. 1442 -1445. Montréal, Canada. Available at: <https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000192.PDF> (accessed September 2019)
- Gravem, D. (2011). Response burden trends and consequences. In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž, March 22-23, 2011. 221-236. Heerlen, The Netherlands: Statistics Netherlands. Available at: <http://www.cbs.nl/NR/rdonlyres/23FD3DF5-6696-4A04-B8EF-1FAACEAD995C/0/2011proceedingsblueets.pdf> (accessed March 2019).
- Haraldsen, G. and Jones, J. (2007). Paper and Web Questionnaires Seen from the Business Respondent's Perspective. In *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, June 18-21, 2007. 1040-1047. Montreal, Canada. Available at: <https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000259.PDF> (accessed September 2019)
- Haas, G.-C., Eckman, S., Bach, R., Kreuter, F. (2016). Is Moving Establishment Surveys from Mail to Web a Good or Bad Decision in Terms of Performance and Data Quality? In *Proceedings of the International Conference for Establishment Surveys 2016 (ICES-V)*. June 21-23, 2016. Geneva, Switzerland. Available at: https://ww2.amstat.org/meetings/ices/2016/proceedings/ICESV_TOC.pdf (accessed February 2020)
- Haraldsen, G., Jones, J., Giesen, D. and Zhang, L. C. (2013). Understanding and coping with response burden. In *Designing and Conducting Business Surveys*, edited by G. Snijders, G. Haraldsen, J. Jones, and D. Willimack, 219 – 252. Hoboken, NJ: John Wiley and Sons.

- Harrell, L., Yu, H. and Rosen, R. (2007). Respondent acceptance of web and E-mail data reporting for an establishment survey. In *Proceedings of the Third International Conference on Establishment Surveys (ICES-III)*, June 18-21, 2007. 1442–1445. Montreal, Canada. Available at: <https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000230.PDF> (accessed September 2019)
- Hedlin, D. (2011). Reducing actual response burden by survey design. In *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*, edited by D. Giesen, 25-32.
- Hedlin, D., Dale, T., Haraldsen, G. and Jones, J. (2005). *Developing Methods for Assessing Perceived Response Burden*. Luxembourg: Eurostat.
- Jones, J, Snijkers, G. and Haraldsen, G. (2013). Surveys and Business Surveys. In *Designing and Conducting Business Surveys*, edited by G. Snijkers, G. Haraldsen, J. Jones and D.K. Willimack, 1-33. Hoboken, NJ: John Wiley & Sons.
- Jones, J. (2012). Response Burden: Introductory Overview Lecture. In *Proceedings of the Fourth International Conference on Establishment Surveys*, June 11-14, 2012. Montréal, Canada. Available at: <http://www.amstat.org/meetings/ices/2012/papers/302289.pdf>. (accessed September 2019)
- Kaplowitz, M. D., Hadlock, T.D. and Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, 68: 94-101. DOI: <https://doi.org/10.1093/poq/nfh006>.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5: 213-236. DOI: <https://doi.org/10.1002/acp.2350050305>.

- Kwak, N. and Radler, B. (2002). A Comparison between Mail and Web Surveys: Response Pattern, Respondent Profile, and Data Quality. *Journal of Official Statistics*, 18: 257-273. DOI: <https://doi.org/10.1177/1525822X08317085>
- Lagerstøm, B. (2018). Chatbots as digital interviewers. *Paper presented at the International Household Nonresponse Workshop*, August 22-24, 2018. Budapest, Hungary.
- Löfgren, T. (2011). Burden reduction by instrument design. In *Response Burden in Official Business Surveys: Measurement and Reduction Practices of National Statistical Institutes*, edited by D. Giesen, 43-50.
- Lyly-Yrjänäinen, M. and van Houten, G. (2011). Reduce burden, increase motivation. Main findings of the quality assessment of the second European company survey. In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen, and M. Bavdaž, 107-118. Heerlen, Netherlands.
- Medway, R. L. and Fulton, J. (2012). When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates. *Public Opinion Quarterly*, 76: 733-746. DOI: <https://doi.org/10.1093/poq/nfs047>.
- Messer, B. L., Edwards, M. L., and Dillman, D. A. (2012). Determinants of Item Nonresponse to Web and Mail Respondents in Three Address-Based Mixed-Mode Surveys of the General Public. *Survey-Practice*, 5: 1-8.
- Millar, M. M., Neill, O. A. C., and Dillman, D. A. (2009). Are Mode Preferences Real? *Technical Report of the Social & Economic Sciences Research Center*. Pullman, Washington: Washington State University.
- Moore, D. and Wojan, T. (2016). From Experimental Testing to Reality: Outcomes from Mixed-mode Implementation of the 2014 National Survey of Business Competitiveness. In *Proceedings of the International Conference for Establishment Surveys 2016*

- (ICES-V). Available at. https://ww2.amstat.org/meetings/ices/2016/proceedings/076_ices15final00063.pdf. (accessed September 2019)
- Rosen, R. and Gomes, T. (2004). Converting CES Reporters from TDE to Web Data Collection. In *Proceedings from the Joint Statistical Meetings*. Toronto, Canada.
- Sear, J. (2011). Response burden measurement and motivation at Statistics Canada. In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž, 151-160.
- Snedecor, G. W. and Cochran, W.G. (1989). *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press.
- Snijkers, G., Onat, E. and Vis-Visschers, R. (2007). The Annual Structural Business Survey: Developing and Testing an Electronic Form. In *Proceedings of the International Conference for Establishment Surveys 2007 (ICES-III)*, 456-463. Montreal, Canada.
- Snijkers, G., Haraldsen, G., Sundvoll, A., Vik, T. and Stax, H. P. (2011). Utilizing web technology in business data collection: Some Norwegian, Dutch and Danish experiences. In *Proceeding of the European Conference on New Techniques and Technologies for Statistics (NTTS)*, Brussels.
- Stang, S. and Thomas, E. (2016). *Web Collection in the Quarterly Census of Employment and Wages Program*. Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.
- Verkruyssen, F. and Moens, S. (2011). Communication as a tool to reduce perceived response burden: Tips and tricks. In *Proceedings of the BLUE-ETS Conference on Burden and Motivation in Official Business Surveys*, edited by D. Giesen and M. Bavdaž, 237-242.
- Willimack, D. and Nichols, E. (2010). A hybrid response process model for business surveys. *Journal of Official Statistics*, 26: 3–24.

- Van Loon, A.J.M., Tijhuis, M., Picavet, H.S., Surtees, P.G. and Ormel, J. (2003). Survey Non-response in the Netherlands: Effects on Prevalence Estimates and Associations. *Ann Epidemiol*, 13: 105-110. DOI: [https://doi.org/10.1016/S1047-2797\(02\)00257-0](https://doi.org/10.1016/S1047-2797(02)00257-0).
- Yan, T., Fricker, S., and Tsai, S. (2020). Response Burden: What Is It and What Predicts It? In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, and A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 193–212). Wiley. <https://doi.org/10.1002/9781119263685.ch8>

Appendix

Table 2.4: Wording and response options for the response burden indicators.

Dimension	Indicator	Question	Response options
Perceived burden	Perception of time	Did you find it quick or time consuming to fill in the questionnaire?	Very quick, Quite quick, Neither quick nor time consuming, Quite time consuming, Very time consuming
	Perception of burden	Did you find it easy or burdensome to fill in the questionnaire?	Very easy, Quite easy, Neither easy nor burdensome, Quite burdensome, Very burdensome
Actual burden	Time to complete (if 1+ persons filled out the questionnaire)	How much time did you spend on actually filling in the questionnaire (sections)?	Number of hours, Number of minutes

Table 2.5: Joint χ^2 values from margin contrast for Minimum Wage and Digitalization questionnaire versions and hypotheses 1-3

	Minimum Wage	Digitalization
<i>H1: lower estimated time for web respondents</i>		
Perceived time	$\chi^2_{4, N=409} = 3.2, p = 0.53$	$\chi^2_{4, N=298} = 0.1, p = 1.0$
Perceived burden	$\chi^2_{4, N=409} = 0.7, p = 0.96$	$\chi^2_{4, N=295} = 0.5, p = 1.0$
<i>H2.1: lower estimated time for choice-paper respondents</i>		
Perceived time	$\chi^2_{4, N=655} = 2.5, p = 0.64$	$\chi^2_{4, N=461} = 0.1, p = 1.0$
Perceived burden	$\chi^2_{4, N=652} = 0.7, p = 0.94$	$\chi^2_{4, N=457} = 0.1, p = 1.0$
<i>H2.2: lower estimated time for choice-web respondents</i>		
Perceived time	$\chi^2_{4, N=305} = 1.2, p = 0.88$	$\chi^2_{4, N=234} = 0.0, p = 1.0$
Perceived burden	$\chi^2_{4, N=304} = 1.5, p = 0.83$	$\chi^2_{4, N=234} = 1.0, p = 0.90$
<i>H3: lower estimated time for choice-web respondents</i>		
Perceived time	$\chi^2_{4, N=551} = 0.66, p = 0.96$	$\chi^2_{4, N=397} = 0.7, p = 0.95$
Perceived burden	$\chi^2_{4, N=547} = 0.0, p = 1.0$	$\chi^2_{4, N=396} = 0.6, p = 0.96$

* $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, and *** $p \leq 0.0$

3 Comparing Single-sitting Versus Modular Text Message Surveys in Egypt

3.1 Abstract

Survey researchers increasingly explore the benefits and drawbacks of text message surveys. This survey mode enables cost efficient data collection and can be applied in hard-to-reach populations where other survey modes suffer from undercoverage, e.g., in regions with low landline and/or low internet penetration. However, so far not much is known about how to best administer surveys in this mode. We experimentally compare two different designs of automated text message surveys in terms of response rate, non-response bias, substantial responses, and participation in a follow-up survey in Egypt. In the *single-sitting* design, respondents automatically received a text message with a new question once they reply to a question. In the *modular* design, respondents received a new question each day, regardless of whether they had responded to the previous question. We invited 1,081 Egyptian parents of kindergarten children who own a mobile phone to participate in a text message survey with eight questions on nutrition behavior of their children. We randomly split the sample into the two design groups. We found that, compared to the single-sitting design, the modular design achieved a higher number of answered questions but had fewer fully completed questionnaires. In addition, we found a difference between groups on substantive responses of behavioral questions. We find no nonresponse bias in both groups and no difference in probability to respond to a follow-up survey. Our results will help researchers making design decisions about how to implement text message surveys.

3.2 Introduction

The use of mobile phones is on the rise worldwide, not only in North America (PEW 2014a) and Europe (GMSA Intelligence 2020), but also in many non-western countries (Silver et al. 2019). While the vast majority of mobile phones in western societies are smartphones, the share of internet-enabled smartphones in emerging countries is still lagging behind (Taylor and Silver 2019). In Egypt, the country of interest for this study, PEW reported a mobile phone penetration rate of 88% while only about 43% of the Egyptian population use the internet (PEW 2014b).

For researchers who want to conduct surveys in these emerging countries, penetration rates have a crucial influence on the decision which mode to use for data collection, and text message surveys may be especially suitable for studying hard-to-reach populations, that is, populations with no permanent addresses, low landline penetration, and low internet penetration. In addition, text messages do not need a permanent network connection making them a viable survey method in regions without a stable phone network connection (Conrad et al. 2017). In fact, a short connection is enough to receive and send multiple messages. Furthermore, text message surveys are considered a very fast mode, i.e., most respondents reply within a day (Conrad et al. 2017, Down and Duke 2003, Hoe and Grunwald 2015, McDonald and Kifer 2018). Another advantage of text message surveys is that they are less intrusive than modes with direct interviewer contact, e.g., face-to-face or telephone interviews (Broich 2015, Johnson 2016). Even if an interviewer conducts the text message survey instead of an automated texting system, text message surveys are considered to be less intrusive (West et al. 2015). At last, text message surveys, especially if conducted without interviewers, are very cheap. For example, Hoe and Grunwald (2015) found that a text message survey only costs a tenth of a telephone study.

While text message surveys offer many advantages, there are also certain limitations researchers should keep in mind when administering text message surveys. For instance, text messages have a character limit depending on country and provider (usually 160 characters per message). Messages longer than this limit are automatically broken up into multiple messages and the order of the messages might change due to technical reasons when participants receive messages. Therefore, researchers are advised to fit a question and all response options in one text message (Conrad et al. 2017, Down and Duke 2003, Johnson 2016). Furthermore, respondents may answer to a survey question by texting back a number or letter associated with the response, the original wording of the response, or some combination or variation of the response. This makes the standardization of answers and the use of plausibility checks much more complicated, and the system used should allow for a variety of valid answers (Down and Duke 2003, Broich 2015).

While researchers are gaining experience with text message surveys, there are still several open questions about how to best implement this mode. For instance, response rates from various text message surveys vary in a wide range from 0.2% to 94.7% (see Table 3.1). So far, no systematic analysis has evaluated which characteristics of text message surveys are responsible for the magnitude of response rates.

Table 3.1: Summary of response rate in text message surveys.

Study	Country	Number of Questions	Response rate in %
West et al. 2015	Nepal	15	94.7
Marlar et al. 2014	US	5 12	13 12
Schober et al. 2015	US	32	48.9 (without interviewer) 71.8 (with interviewer)
Broich 2015			1.3
Down and Duke 2003		2	54
Hoe and Grunwald 2015		5	7
McDonald and Kifer 2018	US	39	32

Study	Country	Number of Questions	Response rate in %
Johnson 2016	Kenia	5 (wave 1)	39.5
		6 (wave 2)	28.4
Lau et al. 2018	Ghana	16	0.6
	Nigeria	16	0.3
	Uganda	16	14.2
	Kenia	16	12.1
Lau et al. 2019	Nigeria	12	0.2
Cooke et al. 2003	UK	3 over 5 days	69
		3 within 1 day	61
		4 within 1 day	58
		5 within 1 day	64

The goal of our study is to contribute to the literature on text message surveys by experimentally comparing two designs how researchers might administer text message surveys. In the *single-sitting* design, respondents receive a survey question via text message and upon answering that question automatically receive the next question. To complete the survey, respondents must answer all questions, like they would do in many other survey modes. In the *modular* design, respondents receive one new question each day, regardless of whether they had responded to the previous question.

Arguments for using the modular design are that such a design might reduce the perceived questionnaire length and thus lowers respondents' burden, limit recall error for questions that refer to behavior on a specific day, and eliminate context effects (West et al. 2015, Johnson et al. 2012, Smith et al. 2012, Peytchev et al. 2020, Toepel and Lugtig 2018). The literature differentiates between two kinds of modularization (see, e.g., Toepel and Lugtig 2018). First, within respondent modularization, meaning a questionnaire is splitted into several modules, and respondents are invited to answer all modules at separate points in time. Survey designers can either control the order in which the modules have to be answered by inviting respondents to each module one after the other, or they can let respondent decide in which order and at what time they want to respond to each module.

Second, modularizing across respondents, meaning that different questionnaire modules are assigned to different groups of respondents. This technique is also known as split questionnaire design.

Our study employs a within respondent modularization with a predefined order in which respondents had to respond and for the sake of simplicity, we will refer to this as the modular design throughout the paper. Our study compares a single-sitting to a modular text message survey design on the following dimensions: unit and item response, non-response bias, substantive responses, and participation in a follow-up survey. While our study provides insights in the use of text message survey methods for a specific population in Egypt, that is, parents who participated in a nutrition health project targeting their kindergarten children, the findings will add to the literature on text message survey design. Our paper is structured as follows. First, we provide an overview about the current state of the literature on text message surveys, including existing research on differences between single-sitting and modular text message survey designs. Second, we describe the nutrition health project in Egypt which our study is part of, and we explain our experimental design. Third, we present the results of our study. Fourth, we provide a discussion of the theoretical and practical implications of our findings as well as suggestions for future research in this field.

3.3 Using a modular design in text message surveys

To formulate hypotheses on how the modularization of a text message survey may affect survey outcomes, we review the current literature of text message surveys and the modular design.

3.3.1 Unit and item response

Two studies compared response rates between single-sitting and modular design in text message surveys. Cooke et al. (2003), who used an automated text message system, conducted an experiment with members of an online panel in Great Britain and found that text message surveys spread out over five days (modular) have a higher response rate than a text message survey administered in a single-sitting in one day. However, they also found that completion rate, i.e., the proportion of respondents who answered all questions, was higher for the single-sitting group compared to the modular group. At first, this effect seems counterintuitive but makes sense if we consider that the modular design works similarly to sending reminders², which are known to increase response rates (e.g., Shih and Fan 2008). However, since the second question replaces the first question in the modular design, respondents who, for instance, missed the first question and started with the second had no possibility to go back and answer the first question and therefore cannot respond to all questions in the survey.

A second study that compared single-sitting and modular designs was conducted by West et al. (2015) in Nepal, who conducted an interviewer administrated text message survey. The authors did not find any differences in response rates between the two text message design groups but they attribute that to the extremely high overall response rate (94.7%) and the extremely cooperative Nepalese respondents. However, the study found that item response was 26 to 53.5 percentage points higher in the single-sitting design group.

Toepel and Lugtig (2018) also conducted an experiment to compare the single-sitting and two modular designs but did so in a web survey. One modular group received three larger survey modules with a gap of one week between each module, and the second received

² Cooke et al. (2003) refrained from using reminders because they had a legitimate response rate in the one day survey.

ten shorter modules with a gap of two days between each module. Respondents were able to participate via PC, smartphone, or tablet. Results from this study are comparable to the text message studies in that the overall number of people who started the survey increased with the number of modules in a design but so did the share of people who dropped out. Eventually, this led to similar number of completes in each of the three groups.

Based on the current literature, we hypothesize that compared to a single-sitting mode, we should see a higher unit response rate (H1) and lower item response rates (H2) for the modular design of a text message survey.

3.3.2 Nonresponse bias

Systematic differences between respondents and the invited sample may lead to biased estimates in substantive variables (Groves et al. 2011). For example, the data collected in our text message survey will be used to produce estimates about health-related behaviors of parents and the health condition of their kindergarten children. If responding to the text message survey correlates with personal or household characteristics that are related to health behaviors, e.g., employment status, income, and education of the parents, resulting estimates may only represent a subset of parents. A large number of studies evaluated sample composition or nonresponse error of text message surveys by comparing characteristics of the respondent sample against the characteristics of the invited sample or another benchmark and find that text message surveys underrepresent women, older people, less educated, less technically savvy people, married people, and people living in rural areas (e.g., Lau et al. 19, Lau et al. 2018, Hoe and Grunwald 2015, Johnson 2016). However, these studies do not evaluate the potential resulting nonresponse bias.

When it comes to nonresponse bias due to the design of the text message survey, West et al. (2015) found no differences in key characteristics between the experimental groups in

their study. However, this result may again be attributed to the very cooperative population. We are not aware of other studies that evaluated nonresponse bias between single-sitting and modular text message survey design. However, the modular design, where respondents can skip a daily question, may work in a similar way as using reminders. Lau et al. (2018) found that the use of reminders in text message surveys increases the proportion of older and less educated respondents, who are usually underrepresented in text-message surveys. As we do not use reminders for the single-sitting group, we may be able to identify different nonresponse bias between single-sitting and modular design in our study. We therefore expect that the respondent sample of the modular design group shows less nonresponse bias than the single-sitting design group (H3).

3.3.3 Substantive responses

The longer a certain event is in the past, the harder it is for people to correctly remember this particular event (see Cannel et al. 1981). Therefore, daily text message surveys that ask about the behavior on that day may produce lower retrospective error compared to a one-time survey that asks retrospective questions about a longer time period. Johansen and Wederkopp (2010) evaluated the use of text message surveys for collecting patients back pain. They compared a weekly text message survey against a one-time retrospective telephone survey asking about respondents' number of days with back pain during the last week³, the last month, and the last year. The authors found no differences in reporting backpain between the text message survey and the telephone survey asking about the last week as well as the monthly aggregate of weekly reports from the text message survey and the telephone survey asking about the last month. However, they found that the reported number of days with back pain largely differed between the aggregated yearly estimate from the weekly text message survey and the one-time report in the telephone

³ Question: "How many days in the past week have you had problems due to LBP?"

survey asking about the last year by 36 days on average, concluding that the weekly text message survey achieves more reliable reports.

West et al. (2015) found no differences between the single-sitting and the modular design concerning substantive responses. However, in this study respondents were not asked about specific events in the past but about behaviors or lifetime histories that are unlikely to change over time. Toepel and Lugtig (2018) found that satisficing behavior decreases with the number of modules in a modular design. Their modules, however, each contain more than one question and their single-sitting design contained over 100 items.

In our study the largest time period for recalling behavior is five days. As Johansen and Wedderkopp (2010) found no difference between the weekly and monthly report, the time differences between our design groups are probably too short to find differences in response behavior that can be tied to recall bias. Therefore, we do not expect any differences between the single-sitting and modular group (H4).

3.3.4 Effects on follow-up survey participation

The main idea of a modular survey design is that splitting the questionnaire in smaller parts reduces the perceived questionnaire length (i.e., respondent burden) and therefore reduces break-offs within a module (Toepel and Lugtig 2018). However, with each additional module, individuals are invited to another short survey. As a result, respondents may feel fatigued and less interested at some point to respond to another survey, and we may see an effect similar to panel attrition. On the other side, a daily invitation to a survey may increase the commitment to the study and the organization conducting the research. Individuals may apply the sunk cost fallacy, that is, they invest more, if they already invested in something (Arkes and Ayton 1999). In terms of modular text message survey

designs, this could mean that individuals that participated in more modules (i.e., invested time), are more likely to respond to future surveys of the same study or from the same organization (i.e., invest more time). As we are not aware of any study that has evaluated the effect of participating in a modular survey on participation in follow-up surveys, we also explored the effect of participating in a modular survey on the response rate of the follow-up survey.

3.4 Data and Methods

To study the effect of the design of text message surveys, we conducted an experiment. Our research is embedded in a larger study evaluating the impact of nutrition of kindergarten children in Egypt in 2017. We refer to this study as the *nutrition study*. The nutrition study targeted 76 kindergartens in Egypt. A baseline face-to-face survey with 3,003 parents of 4,517 children in these kindergartens was conducted in 2017. The baseline survey collected information about parents' and childrens' sociodemographics, childrens' health and nutrition, information on the main caregiver, basic information on the household, and available means of communication in the household.

Among other things, the baseline survey asked the parents whether they own a mobile phone (77.3%). All 2,322 parents reporting to own a mobile phone were asked to provide a phone number and consent to being contacted for a text message survey. To verify that phone numbers were working and a parent could be reached, all phone numbers were called once before sending any text messages. For 1,081 parents (36.0% of all parents in the nutrition study) a working phone number could be verified. This makes up the sample for our text message surveys.

We used the viamo (formerly votomobile) platform to administer two waves of the text message survey. Before the start of the survey, each parent with a valid phone number in

the study received a pre-paid, unconditional incentive as airtime sent to their phone account. This amount would at least cover the costs for sending and receiving the text messages as part of the two survey waves.

The text message survey comprised of two waves (see Figure 3.1). *Wave 1* consisted of eight questions (see Appendix Table 3.7 for question wording and response format) and was administered in a single-sitting design, i.e., respondents were expected to answer the survey in one go. A new question was sent automatically once the previous question was answered, and all questions had to be answered to complete the survey. Therefore, respondents had the chance to stop the survey at any question and return to the questionnaire later. Data collection for wave 1 started on Monday, April 3rd 2017. The survey was left open for a week and no reminders were sent. Overall, 267 parents (24.7%) responded to wave 1 by answering at least the first question, and 102 (9.4 %) parents completed the entire questionnaire by responding to all questions.

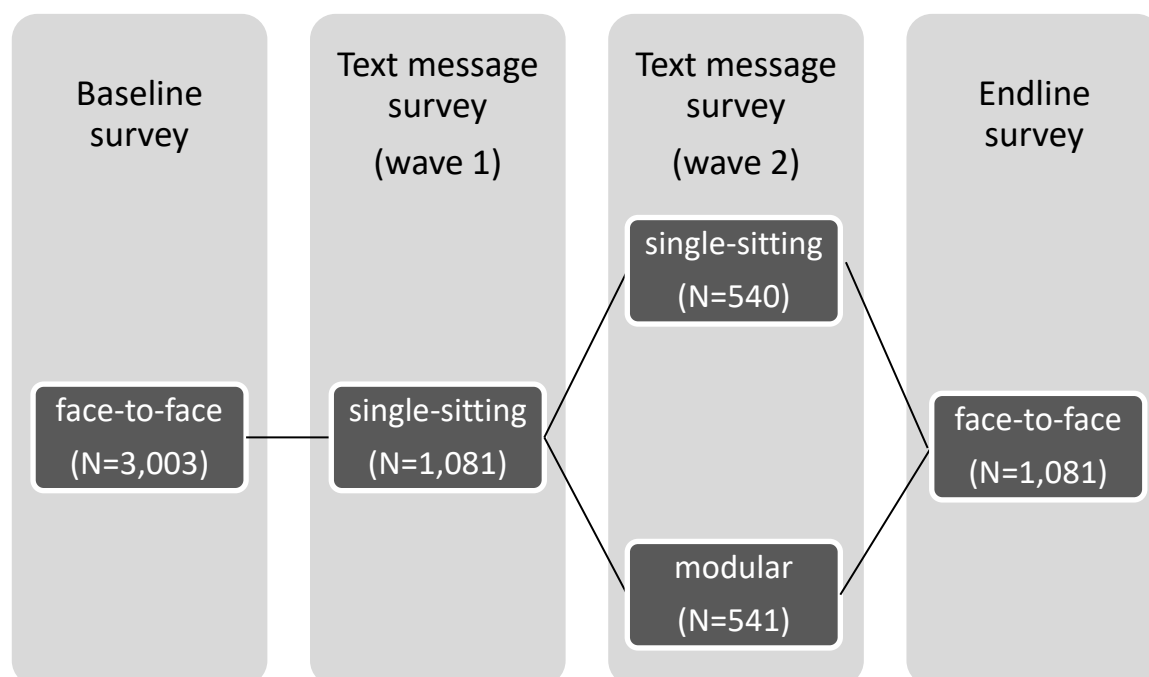


Figure 3.1 Overall design of the nutrition study

Three weeks after the end of wave 1, we invited the same 1,081 parents to participate in the wave 2 survey. This time we conducted a 2 x 2 experiment with random assignment to four groups. The first experiment concerned the two different survey designs for the text message mode: single-sitting vs. modular. The second experiment concerned an intervention on healthy nutrition information⁴.

Table 3.2 shows the field implementation of our experimental design. For the single-sitting group the text message survey design in wave 2 was equal to the one in wave 1: questions had to be answered one after the other, new questions were sent only if a reply to the earlier question was received, and all questions had to be answered to complete the survey. Data collection for the single-sitting group started on Friday, April 28th 2017. The survey was left open for a week, and no reminders were sent. Invited individuals in the modular group received one question each day, starting on Thursday, April 20th 2017. The last question was sent on Thursday, April 27th 2017 (overall eight questions). Parents assigned to the modular group had 24 hours to respond to one question, and once a new question was sent on the next day, they could not go back to the previous question. For both groups, no reminders were sent.

⁴ For the intervention experiment the treatment group (N=541) received a daily text message of helpful tips about healthy nutrition for kindergarten children, aiming at altering parents' awareness and behavior on providing their kids healthy food. The intervention started on Saturday, April 8th 2017 and lasted three days. The control group (N=540) received no text message about healthy nutrition for kindergarten children. We found no evidence that the intervention experiment has any effect on our results. Therefore, we will not discuss results of the intervention experiment.

Table 3.2: Experimental design of text message survey wave 2 for question (Q) 1-8.

Date (month, year, day)	Weekday	Single-sitting	Modular	
April 2017	20	Thursday	-	Q1
	21	Friday	-	Q2
	22	Saturday	-	Q3
	23	Sunday	-	Q4
	24	Monday	-	Q5
	25	Tuesday	-	Q6
	26	Wednesday	-	Q7
	27	Thursday	-	Q8
	28	Friday	Q1-Q8	-

3.4.1 Questionnaire

The questionnaire for text message survey wave 1 and wave 2 included the same eight questions. As an introduction to the text message survey, three messages were sent explaining how to respond to the survey. All messages and survey questions were sent in Arabic. Questions 1 to 3 requested information on parents' nutrition knowledge, and only one answer category could be chosen (see Appendix Table 3.7). Questions 4 to 8 asked parents about the type of food their kindergarten child ate on a specific day. For these questions, respondents could send back all numbers that applied⁵.

The wording of questions four to eight was slightly different in the single-sitting and the modular design (see Table 3.3). In the single-sitting design questionnaire, each question asked retrospectively for the past five days. Since the survey started on a Friday, the questions asked about Sunday, Monday, Tuesday, Wednesday, and Thursday. In the modular

⁵ If respondents choose more answer categories or did not submit a valid number, the question was sent again.

design questionnaire, each daily question asked retrospectively about the previous day, that is, about what their kindergartener ate yesterday. The first of the five questions in the modular group was asked on a Monday. Therefore, the referenced days were equal for each question in the two groups (see Table 3.3).

The answer categories in both groups for each question were the same. Respondents could select multiple answer categories out of five different kinds of foods: “meat or fish”, “fruits”, “vegetables or legumes”, “dairy”, and “sweets”.

Table 3.3: Wording for questions 4 to 8 in the single-sitting and the modular design

Survey design	Question wording	Answer categories (multiple answers allowed)
Single-sitting	On [day]* this week, did your kindergarten kid(s) eat ...?	1 meat or fish 2 fruits 3 vegetables or legumes 4 dairy
Modular	Which of the following did your kindergarten kid eat yesterday ...?	5 sweets

*day: Sunday, Monday, Tuesday, Wednesday, Thursday

3.5 Analysis Plan

3.5.1 Unit and item response

To test H1 (higher unit response rates for the modular design group compared to the single-sitting design group), we compare the proportions of total interviews (answered at least one question), completed interviews (answered all eight questions), partial interviews (answered at least one question but did not finish), and non-participants between

single-sitting and modular design using two sample z-tests of proportions. To test H2 (lower item response rates for the modular design group compared to the single-sitting design group), we compare the proportion of non-missing responses for each question between single-sitting and modular design for all parents who answered at least one question using two sample z-tests of proportions.

3.5.2 Nonresponse bias

To study the influence of the survey design on nonresponse bias, we use information from the baseline survey. Unlike other studies that use person level variables, we use household level variables for our analyses for several reasons. First, data collected in the baseline survey are mostly on household level. Second, variables collected on the person level have low variance due to the homogenous nature of respondents. For instance, respondents in the baseline survey are mainly female (94.7%), indicating that the mother of the child responded to the survey. Third, it is likely that households share a mobile phone. Therefore, it is not certain that the person who responded to the baseline survey also responded to the text message survey. Fourth, health behavior is likely to be homogenous within the household and the quality of nutrition is likely to correlate with factors that indicate wealth, e.g., monthly income and household facilities.

We use the following variables from the baseline survey in our analysis: *employment status of the household breadwinner, monthly income, daily spendings on food, accommodation possessions, household size, number of rooms, and number of daily full meals for children*. We estimate the difference (diff) between the invited sample (b) and the respondent sample (r), i.e., respondents who answered at least one survey question, in a survey design group (d) for each variable (y) the following way:

$$\text{diff}(y_d) = y_{rd} - y_b$$

Next, we calculate the standard error of the estimated bias following the formula used by Lee (2006) and Keusch et al. (under review) as

$$s.e.(y_{rd} - y_b) = \frac{n_b - n_{rd}}{n_b} \times \sqrt{\text{var}(y_{rd}) + \text{var}(y_{nr_d})}$$

To test the significance of a given bias, we use a z-test. To calculate a test statistic, we divide the estimate of the difference by the standard error of the difference.

3.5.3 Substantive responses

To test whether the modular design and the single-sitting design lead to similar or different substantive responses (H4), we compare the answers to the five questions about what food children ate on a specific day. For each referenced day, we compare the proportion of each food category between single-sitting and modular design using Chi-Squared Tests.

3.5.4 Effects on follow-up survey participation

Finally, we specified a logistic regression model to evaluate whether inviting parents to a specific text survey design has an impact on the participation of a face-to-face follow-up survey. The information if a respondent participated in the follow-up survey serves as the dependent variable in our model. The modular design may reduce parents' propensity to participate in three ways. First, by the number of invitations. Therefore, we test if the

group assignment has an impact on the follow-up survey. Second, by responding to at least one question of the text message surveys. Therefore, we include an interaction term between group assignment and the information if an invited parent responded. Third, by the number of questions answered in each group. However, as case numbers are too low (e.g., only four respondents answered all eight questions in the modular group), we cannot consider the number of answered questions as an independent variable for our model. In our model, we control for possible selection effects by including all variables used in our nonresponse bias analysis. For ease of interpretation, we present average marginal effects (AMEs) calculated using the margins package (version 0.3.23) (Leeper 2018) in R version 4.0.3 (R Core Team 2020).

3.6 Results

3.6.1 Unit and item response

Table 3.4 provides an overview about the participation behavior in the single-sitting and the modular design. In the single-sitting design group, 14.3% of contacted parents replied to at least one question with a valid response. In the modular design group, the total response rate is twice as high (28.7%; $p < 0.001$). Therefore, the modular design seems to generate more respondents who complete at least one survey question than the single-sitting design, supporting H1. Similarly, we find a higher rate of partial completes, i.e., respondents who answered at least one question but did not complete the entire survey, in the modular design (27.9%) compared to the single-sitting design (10.4%; $p \leq 0.001$). However, looking at the percentages of completes we see that the single-sitting design produces significantly more completed questionnaires than the modular design (3.9% vs. 0.7%; $p = 0.001$).

Table 3.4: Number and percent of respondents for total, 8 questions, 1-7 questions and each question by experimental design.

	Single-sitting (N=540)		Modular (N=541)		Diff (RR _{modular} – RR _{single-sitting})	
	N	%	N	%	N	%-points
Total Response (responded to at least one question)	77	14.3	155	28.7	78	14.4***
Completes (responded to 8 questions)	21	3.9	4	0.7	-17	-3.2**
Partials (responded to at least one question but did not finish)	56	10.4	151	27.9	95	17.5***
Question 1	77	14.3	43	8.0	-34	-6.3**
Question 2	59	10.9	49	9.1	-10	-1.8
Question 3	49	9.1	52	9.6	3	0.5
Question 4	41	7.6	63	11.6	22	4.0*
Question 5	34	6.3	55	10.2	21	3.9**
Question 6	28	5.2	57	10.5	29	5.3**
Question 7	27	5.0	49	9.1	22	4.1*
Question 8	21	3.9	42	7.8	21	3.9**
Mean number of questions answered ^a	4.9		2.7		2.2***	

p-values are based on two sample z-test of proportions: *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$

^ap-value is based on two sample t-test: *** $p \leq 0.001$

The conceptual differences between single-sitting and modular design also affects the mean number of answered questions. We find that the mean number of answered questions is 4.9 questions for the single-sitting group, which is significantly higher compared to the modular group with an average of 2.7 answered questions ($p < 0.001$). The comparison of mean number of answered questions, suggest a support for H2, that is, we find lower item response rates in the modular design group. However, looking at response on an item level, we see a relatively stable, curvilinear trend in the modular design, i.e., until the fourth question item response rates slightly increase from 8% (Q1) to almost 12% (Q4) and then slightly decrease again to 8% for the last question. For the single-sitting design, we see a steady decrease in the number of responses with each additional question from 14% (Q1) to 4% (Q8). These different trends in the two design groups lead to a significantly lower item response rate for the modular design compared to the single-sitting design in the first question (-6.3 percentage points; $p = 0.001$), while questions 4 to 8 show significantly higher item response rates for the modular design, contradicting H2.

Following our argumentation, Question 1 should have the same response in both groups. However, while respondents in the modular group had 24 hours to respond, the single-sitting group had a few days. Comparing the number of responses for Question 1 and only considering first day responders, we find no statistical difference between groups (10 % vs. 7.6 %; $p = 0.165$).

3.6.2 Nonresponse bias

Table 3.5 shows the nonresponse bias analysis for our baseline variables for both experimental groups. In the following we administer 22 statistical tests for each group; thus we

adjust our p-value with the Bonferroni procedure for multiple comparisons ($p_{Bonferroni\ adjusted} = 0.05 \div 22\ tests = 0.002$).

For both groups, we find no statistically significant ($p > 0.002$) bias for the variables employment status of the household breadwinner, monthly income, daily spendings on food, household size, number of rooms, and number of daily full meals for children. Overall, we only find a few variables that are biased regarding accommodation possessions. In the single-sitting group, we find that the respondent sample has a 2.4 higher percentages points (p.p.) possession rate for refrigerators compared to the invited sample. The possession of a refrigerator, however, is rather high in the invited sample (97.6 %) and the possession rate in the respondent sample is 100 %. It seems likely that the twelve households not owning a refrigerator did not participate by chance and that the difference may be a false positive. In the modular group, we find differences in accommodation possessions between the invited and respondent sample for the accommodation possessions toilet, flush-toilet and hot water. While we find a lower possession rate for toilet (-12.3 p.p., $p < 0.002$), the possession rate for flush toilet (12.3 p.p., $p < 0.00005$) and hot water (11.3 p.p., $p < 0.002$) is higher in the respondent sample of modular design compared to the invited sample. As more variables are biased in the modular group, we see no support for our hypothesis (H3) that the modular group contains less nonresponse bias.

Table 3.5: Comparing nonresponse bias for single-sitting and modular text message survey design by comparing invited with respondent sample for each design

	Single-sitting			Modular		
	Invited sample	Respondents sample	Difference	Invited sample	Respondents sample	Difference
	% (s.e.)	% (s.e.)	Percentage points (s.e.)	% (s.e.)	% (s.e.)	Percentage points (s.e.)
Employment status of household breadwinner						
Unknown status	4.3 (1.7)	2.6 (1.3)	-1.7 (1.8)	5.0 (1.8)	6.5 (2.1)	1.5 (1.6)
Full-time	28.9 (3.8)	36.4 (4.0)	7.5 (5.0)	26.6 (3.7)	31.0 (3.9)	4.4 (3.1)
Part-time	39.6 (4.1)	37.7 (4.1)	-1.9 (5.1)	39.0 (4.1)	35.5 (4.0)	-3.5 (3.3)
Self-employed	22.8 (3.5)	20.8 (3.4)	-2.0 (4.3)	25.0 (3.6)	22.6 (3.5)	-2.4 (2.9)
In fulltime education, unemployed, retired, ill or disabled	4.4 (1.7)	2.6 (1.4)	-1.8 (1.8)	4.4 (1.8)	4.5 (1.8)	0.1 (1.4)
Monthly income						
Unknown	15.4 (3.1)	13.0 (2.9)	-2.4 (3.6)	15.2 (3.1)	9.7 (2.5)	-5.5 (2.2)
1-1000 L.E.	39.6 (4.2)	29.9 (3.9)	-9.8 (4.9)	39.4 (4.2)	37.4 (4.1)	-2.0 (3.3)
1001-2000 L.E.	34.3 (4.1)	49.4 (4.3)	15.1 (5.2)	37.0 (4.1)	41.3 (4.2)	4.3 (3.3)
2001-7000 L.E.	10.7 (2.6)	7.8 (2.3)	-2.9 (2.9)	8.5 (2.4)	11.6 (2.7)	3.1 (2.1)

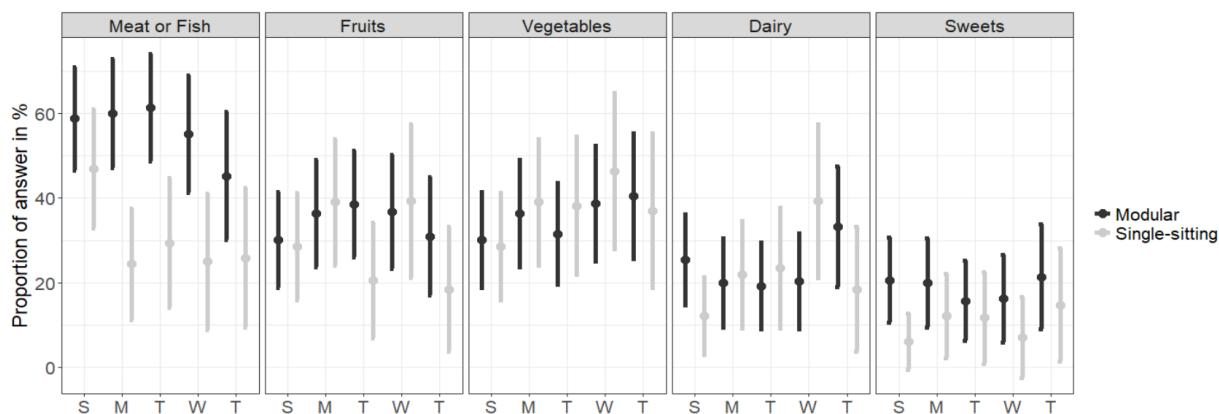
	Single-sitting			Modular		
	Invited sample	Respondents sample	Difference	Invited sample	Respondents sample	Difference
	% (s.e.)	% (s.e.)	Percentage points (s.e.)	% (s.e.)	% (s.e.)	Percentage points (s.e.)
Daily spendings on food						
food is Unknown	6.7 (2.1)	6.5 (2.1)	-0.2 (2.6)	8.9 (2.4)	8.4 (2.3)	-0.5 (1.9)
1-50 L.E.	70.4 (3.8)	71.4 (3.8)	1.0 (4.8)	72.1 (3.8)	69.7 (3.8)	-2.4 (3.1)
51-100 L.E.	18.9 (3.3)	16.9 (3.1)	-2.0 (4.0)	16.1 (3.1)	18.1 (3.2)	2.0 (2.6)
101 or more L.E.	4.1 (1.7)	5.2 (1.9)	1.1 (2.3)	3.0 (1.4)	3.9 (1.6)	0.9 (1.2)
Accommodation possesses						
Toilet	44.6 (4.2)	37.7 (4.1)	-6.9 (5.1)	44.5 (4.2)	32.3 (3.9)	-12.3* (3.2)
Flush-toilet	62.6 (4.1)	68.8 (3.9)	6.2 (4.9)	61.9 (4.1)	74.2 (3.7)	12.3*** (3.1)
Piped water	99.1 (0.8)	98.7 (1.0)	-0.4 (1.2)	98.0 (1.2)	97.4 (1.3)	-0.5 (1.0)
Hot water	53.3 (4.1)	59.7 (4.1)	6.4 (5.2)	55.8 (4.2)	67.1 (3.9)	11.3* (3.2)
Refrigerator	97.6 (1.3)	100.0 (0.0)	2.4** (0.7)	98.2 (1.1)	99.4 (0.7)	1.2 (0.7)
Fan	98.5 (1.0)	97.4 (1.3)	-1.1 (1.6)	98.5 (1.0)	99.4 (0.7)	0.8 (0.7)
Household size ^a	5.5 (0.2)	5.0 (0.2)	-0.5 (2.9)	5.3 (0.2)	5.0 (0.2)	-0.3 (2.4)
Number of rooms ^a	3.1 (0.1)	3.0 (0.1)	-0.1 (1.3)	3.2 (0.1)	3.2 (0.1)	0.0 (1.0)
Number of daily full meals for children ^a	2.6 (0.05)	2.8 (0.06)	0.2 (0.8)	2.6 (0.05)	2.6 (0.05)	0.0 (0.6)

p-values based on z-test: *** $p_{Bonferroni}$ adjusted for $p = .001 < .00005$; ** $p_{Bonferroni}$ adjusted for $p = .01 < .0005$; * $p_{Bonferroni}$ adjusted for $p = .05 < .002$

^a Estimates provided as means not percentages

3.6.3 Substantive responses

Figure 3.2 shows the proportion of respondents who reported that their kindergartner ate a specific food category on a day by text message survey design. Contrary to our hypothesis H4, parents in the modular group more often choose the “meat or fish” category on all five days compared to parents in the single-sitting design group, with non-overlapping confidence intervals on two days (Monday and Tuesday). For the category “sweets”, we see a similar relationship: respondents in the modular design group more often reported that their kindergartner ate sweets compared to the single-sitting design group. However, due to the relatively small sample size, all confidence intervals overlapped for this food category. For the categories “fruits”, “vegetables”, and “dairy” there are no differences in substantive responses between the two groups.



Note: The number of cases changes by days and survey design respectively: $N_{\text{Sunday, modular}} = 63$, $N_{\text{Sunday, single-sitting}} = 49$, $N_{\text{Monday, modular}} = 55$, $N_{\text{Monday, single-sitting}} = 41$, $N_{\text{Tuesday, modular}} = 57$, $N_{\text{Tuesday, single-sitting}} = 34$, $N_{\text{Wednesday, modular}} = 49$, $N_{\text{Wednesday, single-sitting}} = 28$, $N_{\text{Thursday, modular}} = 42$, $N_{\text{Thursday, single-sitting}} = 27$

Figure 3.2: Proportion (points) of food entries for the nutrition question by day of the week Sunday (S) – Thursday (T) with 95% confidence intervals (lines).

3.6.4 Effects on follow-up survey participation

Table 3.6 shows four different models that help us to estimate the effect of the text message survey design (single-sitting vs. modular) on the propensity to participate in a follow-up face-to-face survey. Model 1 shows that there is no main effect of the text message survey design on the propensity to response to the follow-up survey. Controlling for possible selection effects by adding household level characteristics in Model 2 does not change the result. While this is not one of our main research questions, it is interesting to note that households with a monthly income between 2001-7000 L.E. have a significantly lower probability to participate in the follow-up survey than parents with an income between 1-1000 L.E. The probability to participate in the follow-up face-to-face survey slightly increases with the number of rooms in the home and with spending 51-100 L.E. on food daily compared to spending 1-50 L.E. on food daily.

Table 3.6: Average marginal effects (AME) for logistic regression models to evaluate the impact of participation on follow-up surveys

	Model 1	Model 2	Model 3	Model 4
	AME (s.e.)	AME (s.e.)	AME (s.e.)	AME (s.e.)
Survey design (ref. = modular)				
Single-sitting	-0.02 (0.03)	-0.02 (0.03)	-0.04 (0.03)	-0.03 (0.03)
Responded to at least one question in wave 1 (ref. = did not respond)				
Responded			-0.01 (0.04)	-0.01 (0.03)
Responded to at least one question in wave 2 (ref. = did not respond)				
Responded			-0.09 (0.04)*	-0.07 (0.04)
Interaction between survey design and response to text message survey wave 2 (ref. did not respond)				
Response to text message survey wave 2 in single-sitting			-0.09 (0.06)	-0.07 (0.06)
Response to text message survey wave 2 in modular			-0.09 (0.05)	-0.07 (0.05)
Employment status of household breadwinner (ref. = Part-time)				
Full-time		-0.02 (0.04)		-0.02 (0.04)
Self-employed		-0.01(0.04)		-0.01 (0.04)
In fulltime education/ Unemployed/ Retired/ Ill or disabled		-0.08 (0.08)		-0.09 (0.08)
Unknown status		-0.02 (0.07)		-0.01 (0.07)
Monthly income (ref. = 1-1000 L.E.)				
1001-2000 L.E.		-0.01 (0.03)		-0.01 (0.03)
2001-7000 L.E.		-0.18 (0.06)**		-0.19 (0.06)**

	Model 1	Model 2	Model 3	Model 4
	AME (s.e.)	AME (s.e.)	AME (s.e.)	AME (s.e.)
Unknown		-0.02 (0.05)		-0.03 (0.05)
Daily spendings on food (ref. = 1-50 L.E.)				
51-100 L.E.		0.08 (0.04)*		0.08 (0.04)*
101 or more L.E.		-0.11 (0.09)		-0.11 (0.09)
Unknown		-0.01 (0.06)		-0.01 (0.06)
Accommodation possesses				
Toilet		0.0 (0.05)		0.0 (0.05)
Flush-toilet		0.02 (0.06)		0.02 (0.06)
Piped water		0.03 (0.12)		0.02 (0.12)
Hot water		-0.04 (0.04)		-0.03 (0.03)
Refrigerator		-0.2 (0.13)		-0.18 (0.13)
Fan		0.13 (0.12)		0.13 (0.12)
Household size		0.01 (0.01)		0.01 (0.01)
Number of rooms		0.03 (0.01)*		0.03 (0.1)*
Number of daily full meals for children		-0.03 (0.02)		0.03 (0.01)
N	1,081	1,081	1,081	1,081
AIC	1,405.2	1,408.4	1,404.6	1,409.8

Model 3 and 4 in Table 3.6 include indicators for whether a parent had responded to the text message survey in wave 1 and whether they had participated in the text message survey in wave 2 together with an interaction term for text message survey participation in wave 2 and the design of the survey in wave 2. None of the coefficients is statistically significant, indicating that responding to the text message survey had no effect on the probability to participate in the follow-up survey.

3.7 Conclusion

In this article, we investigated the effects of two designs to administer text message surveys: single-sitting and modular. We randomly assigned 1,081 Egyptian parents of kindergarten children who participated in a nutrition study to the group *single-sitting*, in which parents received one invitation to an eight question long text message survey and to the group *modular* in which parents received an invitation to a question each day over the course of eight days. We evaluated the effect of both groups on unit and item response rates, substantive responses and effect on response propensity on a follow-up face-to-face survey.

We found that the modular group is able to recruit more respondents than the single-sitting group. However, the increase in the number of respondents goes at the expense of the average number of questions answered by the respondent. Furthermore, we see an interesting trend between both groups. While the item response rates decrease from the first (14.3 %) to the last question (3.9 %), item response rates are distributed curvilinear ranging from 7.8 to 11.6 %. Reason for this distribution may relate to the fact that respondents in the single-sitting group have to answer the prior question to proceed within the survey, i.e., to answer the second question, the first question must be answered; to

answer the fifth question, the fourth question must be answered. Therefore, we see a decreasing number of respondents with each question. For the modular group, however, invited individuals can decide each day if they want to continue with the survey, regardless of whether they answered the question on the previous day or not. Against this background, the modular design may work in a similar way as a reminder. Therefore, using a reminder for parents in the single-sitting design may increase the response rate and decrease the difference in response rates between single-sitting and modular.

Our second hypothesis, that is, lower item response rates in the modular design group compared to the single sitting group, is not supported by our analysis. This hypothesis was mainly based on a text message study in Nepal (West et al. 2015) which found that item response was 26 to 53.5 percentage points higher in the single sitting design group. The study had very cooperative respondents with an overall response rate of 94.7 % that did not differ between experimental groups. We assume that the difference between West et al. (2015) and our study are due to different target populations.

Using a reminder in text message surveys has a positive effect on sample composition (Lau et al. 2018). The modular design, i.e., sending a survey invitation each day, may have a similar effect as sending reminders and may have a positive effect on nonresponse bias. Therefore, we hypothesized that the respondent sample in the modular group should have lower nonresponse bias than the respondent sample of the single-sitting group in which respondents were only invited once. We find no support for this hypothesis. However, we find that compared to the invited sample, respondents in the modular groups are more likely to have a flush toilet and a hot water supply. Both accommodation possessions may be associated with a higher wealth. However, more direct wealth indicators like monthly income or daily spending for food are unaffected. Therefore, we cannot

conclude that the modular design increases the propensity to respond for more wealthier households.

Five of the text message survey questions were retrospective behavioral questions that asked parents which kind of food they provided to their children on a certain day. While in the single-sitting group those questions related to the last five days, in the modular group those questions related to the day before the survey invitation to each question. Parents could choose between five answer categories “meat or fish”, “vegetables”, “fruits”, “dairy” and “sweets”. Our data shows that parents in the modular group are more likely to state that their kids ate “meat or fish” and “sweets”. Providing meat and fish to children as a healthy nutrition may be seen as a social desirable behavior. On the other side, giving sweets to children is a less social desirable behavior in terms of healthy nutrition. We are not able to explain the underlying effects behind those differences. Nevertheless, we find some differences in responses between the modular and single-sitting design.

For the effect on follow-up surveys, we were not able to identify an effect that suggests that inviting parents multiple times and parents responding in the modular group have a lower propensity to participate in a follow-up survey.

Finally, we like to discuss some limitations of the study and using text message surveys as a modular design approach.

First, our modularization had one question for each module which may be not efficient. For instance, respondents may perceive the participation of one question as not rewarding and may expect more and feel more burdened by participating each day in a survey instead of answering a survey in one sitting.

Second, in the modular design, respondents can only answer the question to which they were invited last. If respondents missed to answer a question yesterday ($n-1$) and already

received a new question today (n), it is not possible to answer yesterday's question as each response will be matched to the today's question. Against this background, respondents may have answered more questions in the modular design on the second or following days, but our data collection design did not notice or replace responses.

Third, we found entries in the modular group with a timestamp indicating that the question was answered the following day but before the invitation to the next module. As each question in the modular group has the wording "yesterday" in it, referencing the previous day, it is not clear if parents answered the question for yesterday or the day we intended: the day before yesterday.

Fourth, as parents in the modular group are invited each day to answer a survey, each module of the modular text message survey can be seen as an independent survey. From this perspective the modular design would not have item level response but unit level nonresponse for each module. Therefore, nonresponse analyses on the item level would be more reasonable. However, we are not able to conduct a sample composition analysis on item level between single sitting and modular as we do not have enough cases for later questions in the single-sitting group.

Our study suffers from a few limitations. However, our results provide important insights and contribute knowledge to the literature on text message surveys by experimentally comparing two designs how researchers might administer text message surveys and will help researchers to make design decisions on how to implement text message surveys.

References

- Arkes, H. R., and Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, 125(5), 591–600.
<https://doi.org/10.1037/0033-2909.125.5.591>
- Broich, C. (2015). Offline data collection in Sub-Saharan Africa using SMS surveys: lessons learned. *Paper presented at the meeting of the American Association for Public Opinion Research*, Hollywood, FL.
- Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on Interviewing Techniques. In S. Leinhardt (Eds.): *Sociological methodology*: 389-437. San Francisco: Jossey-Bass.
- Cooke, M., Nielsen, A., and Strong, C. (2003). The use of SMS as a research tool. In *Proceedings of the fourth ASC international conference* edited by R. Banks, J. Currall, J. Francis, L. Gerrard, R. Khan, T. Macer, M. Rigg, E. Ross, S. Taylor, and A. Westlake. 267–276. Chesham: Association for Survey Computing.
- Conrad, F. G., Schober, M. F., Antoun, C., Hupp, A. L., and Yan, H. Y. (2017). Text interviews on mobile devices. *Total survey error in practice*, 299-318.
<https://doi.org/10.1002/9781119041702.ch14>
- Down, J. and Duke, S. (2003). SMS polling. A methodological review. In *Proceedings of the fourth ASC international conference* edited by R. Banks, J. Currall, J. Francis, L. Gerrard, R. Khan, T. Macer, M. Rigg, E. Ross, S. Taylor, and A. Westlake. 277–286. Chesham: Association for Survey Computing.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey Methodology* (2nd ed.). Hoboken: John Wiley & Sons.

Retrieved from <http://gbv.eblib.com/patron/FullRecord.aspx?p=819140>.

GMSA Intelligence. (2020). The Mobile Economy Europe 2018. Available at:

https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Europe.pdf (accessed February 2021).

Hoe, N.D. and Grunwald, H.E. (2015). The role of automated SMS text messaging in survey research. *Survey Practice*, 8(6).

Johansen, B., and Wedderkopp, N. (2010). Comparison between data obtained through real-time data capture by SMS and a retrospective telephone interview. *Chiropractic & Osteopathy*, 18, 7p. <https://doi.org/10.1186/1746-1340-18-10>

Johnson, D. (2016). Collecting Data from mHealth Users via SMS Surveys: A Case Study in Kenya. *Survey Practice* 9(1). <https://doi.org/10.29115/SP-2016-0004>.

Johnson, A., Kelly, F., and Stevens, S. (2012). Modular survey designs for mobile devices. A research study on the potential uses of modular survey design for mobile and online surveys. *Presented at the CASRO Online Research Conference Las Vegas*, March 1st and 2nd, 2012.

Keusch, F. Bähr, S. Haas, G., Kreuter, F. and Trappmann, M. (under review). Nonparticipation in Smartphone Data Collection Using Research Apps. *Journal of the Royal Statistical Society: Series A*.

- Lau, C.Q., Lombaard, A., Baker, M., Eyerman, J., Thalji, L. (2018). How Representative Are SMS Surveys in Africa? Experimental Evidence From Four Countries. *International Journal of Public Opinion Research*, 30(2).
<https://doi.org/10.1093/ijpor/edy008>.
- Lau, C. Q., Cronberg, A., Marks, L., and Amaya, A. (2019). In Search of the Optimal Mode for Mobile Phone Surveys in Developing Countries. A Comparison of IVR, SMS, and CATI in Nigeria. *Survey Research Methods*, 13(3), 305-318.
<https://doi.org/10.18148/srm/2019.v13i3.7375>.
- Lee, S. (2006). An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey. *Social Science Computer Review*, 24(4), 460–475.
<https://doi.org/10.1177/0894439306288085>.
- Leeper, T.J. (2018). margins: Marginal Effects for Model Objects. R package version 0.3.23.
- Marlar, J., McGeeney, K., and Chattopadhyay, M. (2014). SMS surveys: Testing multiple modes to reach respondents from a wireless frame. *Paper presented at the 69th Annual Conference of the American Association for Public Opinion Research*, Anaheim, CA, May 15–18.
- McDonald, B., and Kifer, M. J. (2018). Using SMS in Mobile Data Collection - Recruitment, Cost & Response. *Paper presented at the 73rd Annual Conference of the American Association for Public Opinion Research*, Denver, CO.
- Pew Research Center. (2014a). Mobile technology fact sheet. Available at:
www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet/ (accessed February 2021).

- Pew Research Center. (2014b). Emerging Nations Embrace Internet, Mobile Technology. Available at: <https://www.pewresearch.org/global/2014/02/13/emerging-nations-embrace-internet-mobile-technology/> (accessed February 2021).
- Peytchev, A., Peytcheva, E., Conzelmann, J.G., Wilson, A. and Wine, J. (2020). Modular Survey Design: Experimental Manipulation of Survey Length and Monetary Incentive Structure, *Journal of Survey Statistics and Methodology*, 8(2), Pages 370–384, <https://doi.org/10.1093/jssam/smz006>.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Schober, M.F., Conrad, F.G., Antoun, C., Ehlen, P., Fail, S., Hupp, A.L., Johnston, M., Vickers, L., Yan, H., and Zhang, C. (2015). Precision and disclosure in text and voice interviews on smartphones. *PLoS One*, 10(6). <https://doi.org/10.1371/journal.pone.0128337>.
- Shih TH, Fan X (2008) Comparing response rates from Web and mail surveys: a meta-analysis. *Field Methods* 20: 249–271.
- Silver, L., Vogels, E.A., Mordecai, M., Cha, J. Rasmussen, R. and Rainie, L. (2019). Mobile divides in Emerging Economies. Available at: https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/11/PG_2019.11.20_Mobile-Divides-Emerging-Economies_FINAL.pdf (accessed February 2021).
- Smith, R., Kotzev, I., Millar, C. and Kachhi, D. (2012). Modularizing surveys to meet today's respondent challenges in 2012 CASRO Online Conference. Available at: <https://c.ymcdn.com/sites/www.casro.org/resource/collection/E270CC91-6B72->

[4C37-BCC0-5503CBB66C55/Paper - Renee Smith and Chuck Miller - Kantar and uSamp.pdf](#) (accessed February 2021).

Taylor, K. and Silver, L. (2019). Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally. Available at: https://www.pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf (accessed February 2021).

Toepoel V. and Lugtig P. (2018). Modularization in an Era of Mobile Web: Investigating the Effects of Cutting a Survey Into Smaller Pieces on Data Quality. *Social Science Computer Review*. <https://doi.org/10.1177/0894439318784882>.

West, B. T., Ghimire, D. and Axinn, W. G. (2015). Evaluating a Modular Design Approach to Collecting Survey Data Using Text Messages. *Survey Research Methods*, 9(2), 111–123. <https://doi.org/10.18148/srm/2015.v9i2.6135>.

Appendix

Table 3.7: Questionnaire wording for text message surveys (wave 1, wave 2 single-sitting and wave 2 modular)

Wave 1 and wave 2 single-sitting (Q sent after response to previous Q is received)	Wave 2 modular (one Q per day)*
<u>Message 1:</u> [155 characters] Thanks for participating in our study. In the next days, you will receive questions via SMS about what your kindergarten kid(s) eat. <u>Message 2:</u> [137 characters] If at any point you don't want to participate in the study anymore, just send an SMS with the word STOP and you will receive no more SMS. <u>Message 3:</u> [161 characters] First, we will ask you what you think about different foods. Just reply with the number that you think is the best answer. The costs of replying are covered by the amount that was sent to your phone. <u>Q1:</u> [133 characters] To be in good health, how often should a kid eat fruits and vegetables? 1 To every meal 2 Every day 3 Every week 4 Less than every week	
<u>Q2:</u> [99 characters] Which food provides your kid with a high amount of iron? 1 Potatoes 2 Nuts and seeds 3 Bread 4 Cake	

<u>Q3:</u> [90 characters] Which of the following best helps your kid to absorb iron? 1 Milk 2 Lemon 3 Bread 4 Sweets	
<u>Q4 - Q8:</u> [154 characters] On [weekday]* this week, did your kindergarten kid(s) eat: 1 meat or fish 2 fruits 3 vegetables or legumes 4 dairy 5 sweets? Send back all numbers that apply. *weekday: Q4 – Sunday, Q5 – Monday, Q6 – Tuesday, Q7 – Thursday, Q8 – Friday	<u>Q4 - Q8:</u> [163 characters] Which of the following did your kindergarten kid eat yesterday? 1 meat or fish 2 fruits 3 vegetables or legumes 4 dairy 5 sweets? Send back all numbers that apply.

4 Effects of Incentives in Smartphone Data Collection

4.1 Abstract

Smartphones are increasingly attractive data collection devices. In particular, they allow us to collect sensor data and analyze phenomena that cannot be investigated with survey data alone. Sensor data collected on smartphones include very sensitive information, such as geolocation and mobility data or app usage data that may be perceived as too private to share with researchers. Therefore, sensor data may be more valuable than survey data to participants, and it may be harder to recruit participants for an app study involving smartphone sensor data than for a survey. However, respondent burden may be reduced by this passive kind of data collection, which might make it easier to recruit participants. In surveys, monetary incentives are known to increase response rates. However, to date, we do not know whether incentives work the same way in studies involving smartphone sensor data. This paper reports results of an experimental study conducted in Germany, in which different incentive amounts and different incentive schemes (paid using Amazon.de vouchers) were randomly assigned to Android users selected from a large labor market panel survey (Panel Study Labour Market and Social Security – PASS). We find that a higher installation incentive resulted in a higher installation rate, but we find very little effect heterogeneity within the experimental conditions and no interaction effects of incentive schemes and wealth.

4.2 Introduction

Smartphone sensor data enable researchers to analyze phenomena that cannot be investigated with survey data alone (e.g., Sugie 2018). However, smartphone data may include very sensitive information, e.g., on geolocation or app usage, which users may perceive as too private to share with researchers. To date, very little research has systematically examined participation in studies that collect passive smartphone sensor data and to our

knowledge, and no study so far has examined whether the knowledge about the effectiveness of incentives in surveys also holds for smartphone sensor data collection. It is possible that common incentive amounts and incentive schemes traditionally used in surveys will not be sufficient to motivate participation in a study that collects data passively from smartphones, given that individuals may perceive sensor data as being more valuable than survey data. However, since participation in passive data collection requires less effort from participants, burden (measured in time spent on data collection) is much reduced compared to regular surveys. Therefore, it might be much easier to recruit participants, and the effect of incentives on participation might be less pronounced. In either case, it is important for researchers to know whether vulnerable groups are particularly receptive to incentives, compared to majority groups in a population. Institutional review boards and ethics committees would likely hesitate to approve a study that—by using monetary incentives—places vulnerable populations, e.g., welfare recipients, at greater risk of providing sensitive data. For these reasons, we not only analyze effects of different incentive schemes on participation rates in a study combining self-reports and passive data collection using smartphones but also break out these effects by economic subgroups. In section 4.3, we start with a brief review of the literature on the effectiveness of incentives and the postulated mechanisms explaining these effects. Section 4.4 explains the study design with an emphasis on the experimental conditions (more details on study design features are described in Kreuter et al. 2018). Section 4.5 displays the results, which we will discuss in section 4.6 paired with suggestions for future research.

4.3 The Influence of Incentives on Participation

Providing some form of incentive, whether monetary or some other kind of token of appreciation, is common for studies recruiting respondents to answer survey questions (see,

for example, James and Bolstein 1990, Church 1993, Willimack et al. 1995, Singer et al. 1999, Singer 2002, Toepoel 2012, Pforr 2016). Singer and Ye (2013) summarize the findings of two decades of research on this topic and state that monetary (cash) incentives are more effective in increasing response rates than are gifts or in-kind incentives, and prepaid incentives are more effective than are promised incentives. Although incentive amounts have increased over time, Singer and Ye (2013:18) also report that research points to the nonlinear effect of monetary incentives, though generally higher incentives increase response rates more than lower ones do.

To gauge how findings from surveys translate to data collection on smartphones, it is helpful to remind ourselves about the different mechanisms suggested to explain the effect of incentives. Going all the way back to the 1960s, Singer and Ye (2013: 115) point to social exchange and the “norm of reciprocity” as an explanation for the effectiveness of prepaid incentives. Social exchange theory argues that prepaid incentives create an obligation to provide on individuals, which they can settle by responding to the survey. The effectiveness of promised incentives—paid conditionally after the survey has been completed—could be better explained by various “versions of utility theories” (Singer and Ye 2013: 115), arguing that people decide on a course of action if, in their view, the benefits of acting outweigh the costs. Since we only use promised incentives, we can only test the hypotheses of the utility theories framework—albeit, as we will explain below, we paid out incentives continuously and respondents did not have to wait until the whole study was over.

In the context of utility theories, the question arises of what value a given incentive has for an individual. In their discussion of Leverage-saliency Theory, Groves et al. (2000) emphasize the relative importance of various features of the survey in the decision-making process, together with how salient these features are to the sample case. Incentives

can be used as leverage to increase survey participation, and Groves et al. (2006) demonstrated that people with low interest in a survey topic can be recruited by monetary incentives that compensate for the lack of interest. Therefore, higher incentives may be used to increase the leverage, whereby research suggests that, compared to lower incentives, higher incentives have a diminishing marginal utility to increase response rates (Singer and Ye 2013, Mercer et al. 2015).

Experiments in panel studies suggest that incentives also have a long-term effect on survey participation. Incentives only need to be paid in one wave to increase participation for the current and following waves, and larger incentives lead to higher response rates in later waves (Singer and Kulka, 2002; Goldenberg, McGrath, and Tan, 2009). For continuous data collection in a smartphone app study, this result could mean that a higher incentive for installing the app may increase participant's commitment throughout the data collection period to keep the app installed and lower attrition or that a higher incentive for installing the app may nudge participants to allow more passive data permissions.

Jäckle et al. (2019) are the first to evaluate the effect of incentives on installing a research app. The app served as a data collection instrument for a spending study in the United Kingdom. Participants had to download the app and upload receipts over the course of a month. The authors randomly assigned sample members into two groups and offered £2 or £6 for downloading the app. For each of the three examined outcomes (completion of the registration survey, proportion of individuals using the app at least once, and proportion of individuals using the app at least once per week over the data collection period), the £6 incentive produced higher rates than did the £2 incentive, but the differences were not statistically significant for any of the outcome variables. The lack of effect might have been a result of the small monetary difference between the two incentive groups.

A concern often voiced in the context of incentives is that the same monetary amount has a higher value for individuals with less wealth (Philipson, 1997, Felderer et al. 2018). If this observation is indeed true, economically disadvantaged sample units might be more inclined to provide data in general and sensitive sensor data in particular.

If incentive payments are perceived as compensation for the time and effort a respondent provides (Philipson, 1997), the opportunity cost for low-income respondents should be lower compared to high-income groups, and incentives should have a stronger effect on low-income respondents. This notion is supported by findings from Mack et al. (1998), who found that a \$20 incentive, compared to a \$10 incentive and to no incentive, proportionally increased participation of respondents with less wealth. Singer et al. (1999) also found that incentives increased response propensities of low-income individuals.

In the context of a smartphone app study such as ours described below, it is not clear whether utility theories are applicable in a similar fashion. Compared to telephone and face-to-face interviews, relatively little time is needed to install an app and to have it run in the background for data collection. However, one could argue that in the case of research apps, it is not time that is exchanged for money but data. In general, if asked hypothetically, individuals are concerned about their privacy when asked to share their data passively with researchers (Jäckle et al. 2019, Revilla et al. 2018, Keusch et al. 2017, Wenz et al. 2017). Those concerns may be tied to trust issues, meaning that individuals do not trust researchers to protect their data adequately. However, individuals' concerns seem to decrease, i.e., their willingness to participate increases, if they are offered more control over when the data are collected, if the study is sponsored by a university (compared with a governmental institution), and if the study offers incentives (Keusch et al. 2017). Cantor et al. (2008) and Dillman et al. (2014) point out that incentives can be used

to establish trust and that trust is more important for gaining cooperation than the incentive value. If these observations describe the major mechanism, we would expect a certain threshold to be needed to establish trust but would not necessarily expect increasing incentive amounts to have a linear effect on participation.

The study presented below has several characteristics that are novel with respect to issuing incentives. First, incentives are paid for installing an app that passively collects data (if the participant grants informed consent) and presents short surveys to participants. Second, incentives are paid for the actual permission to collect such passive data for 30 consecutive days. Without this permission, the app only collects survey and para data (i.e., time stamps and information on whether the data sharing is activated). Third, incentives are paid for answering survey questions. For the first two tasks, the amount and conditions of the incentives were randomly varied, and it is the effects of these variations on participation behavior we examine below in detail.

Broadly speaking, we try to answer the following questions: Do we observe effects of different incentive amounts on the installation rate of a research app? Are these effects proportional to the incentive amounts provided? Do incentives affect participants' decisions to share passive data or to deinstall the app? Do we observe differential effects of incentives, with vulnerable (less wealthy) groups being more receptive to higher incentives than non-vulnerable groups? How much money is actually paid out, i.e., participants downloading vouchers from the app?

4.4 IAB-SMART study design

The goal of the IAB-SMART study is to gain insights into the effects of long-term unemployment on social inclusion and social integration and to examine effects of network integration on reintegration into the labor market using a new data collection approach.

There is not enough room here to go into detail about the planned measurement, but for context, three sets of data are needed to achieve the substantive research goals: (1) a reliable data source about the employment status of the study participants (available at the IAB through social security administration records; see Jacobebbinghaus and Seth 2007), (2) background variables on the study participants (available through IAB surveys), and (3) behavioral measures on the amount of social interaction and network activities (available through the IAB-SMART study). For more details on the overall study design and measurements, see also Kreuter et al. (2018).

4.4.1 Sampling frame and sample restrictions

Participants for the IAB-SMART study were sampled from the German panel study “Labour Market and Social Security” (PASS), an annual household panel survey of the German residential population aged 15 and up oversampling households receiving welfare benefits. PASS is primarily designed as a data source for research into the labor market, poverty, and the welfare state. However, PASS also focuses on the social consequences of poverty and unemployment, including social exclusion and health outcomes. At the time of the IAB-SMART study, PASS has been in the field for 12 years (more information on PASS can be found in the yearly PASS methods and data reports available at https://fdz.iab.de/de/FDZ_Individual_Data/PASS.aspx). Due to the ability to match data collection outcomes of PASS against high-quality administrative records from social security notification and labor market programs, extensive nonresponse studies are available for PASS, showing rather small biases for a range of variables such as benefit receipt, employment status, income, age, and disability (Kreuter et al. 2010, Levenstein 2010, Sakshaug and Kreuter 2012). Foreign nationals have been found to be considerably underrepresented in PASS (Kreuter et al. 2010), but weighting can adjust for this underrepresentation.

All PASS respondents who participated in wave 11 (2017) and reported having an Android smartphone (see Figure 4.1) were eligible for the IAB-SMART app study. We restricted the study to Android devices because extensive passive data collection is restricted under iOS (Harari et al. 2016), and other operating systems had too low market shares to justify additional programming efforts. For the purposes of the incentive study, we do not expect the operating system to have any limiting factor, though we will come back to this point in the discussion section. Keusch et al. (2018) examined issues of coverage and found that smartphone owners in Germany are younger, more educated, and more likely to live in larger communities than are non-smartphone owners, but the authors reported little coverage bias in substantial PASS variables due to smartphone ownership. These results hold even when limiting the sample to Android smartphone owners only.

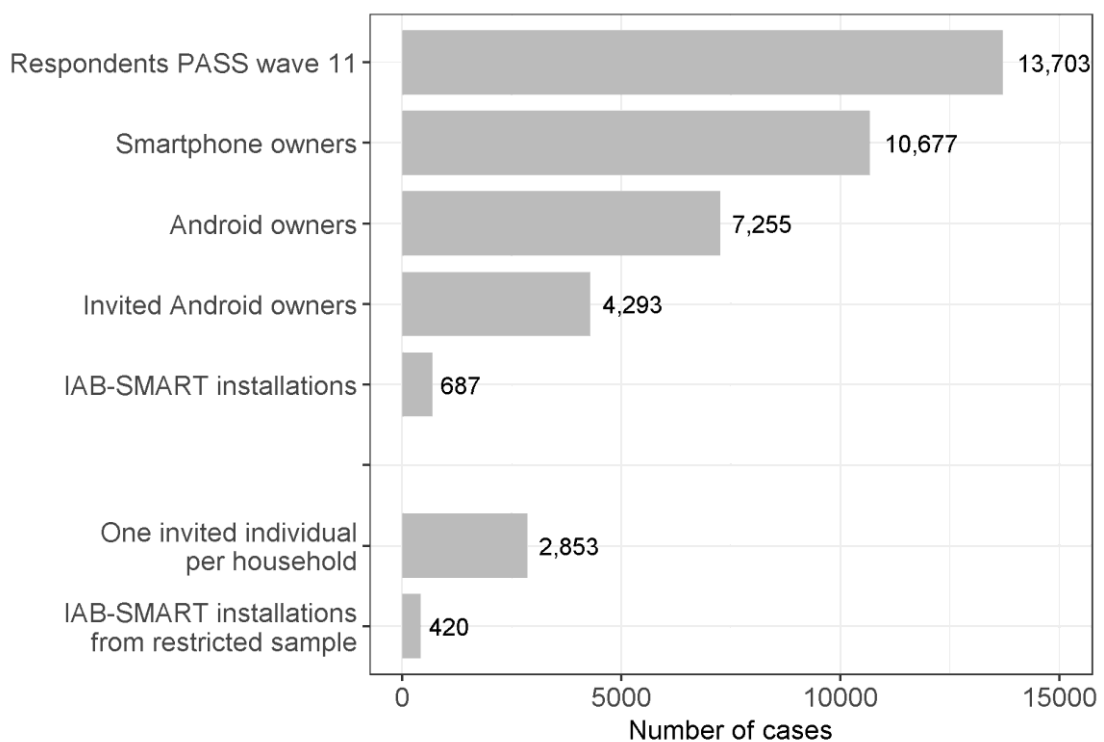


Figure 4.1 Sample size at each stage of the IAB-SMART selection process

4.4.2 Invitation and data request

The invitation to participate in the IAB-SMART study was sent out in January of 2018 to a random sample of PASS Wave 11 participants aged 18 to 65 who reported owning an Android smartphone (N=4,293). To participate in the study, smartphone owners needed to install an app on their smartphone and activate the data sharing functions explained below. The field period of the study was six months.

Our initial goal was to recruit 500 participants. Because response rates to smartphone data collection with extensive data sharing functionalities were hard to gauge from the literature, we sent out invitations in two installments. We used the first round of invitations to 1,074 PASS participants to monitor uptake rates. The second round of invitations was sent to an additional 3,219 PASS participants.

The invitation package sent in both installments contained several pieces: a cover letter (explaining the goals of the study and how to find the app in the Google Play store), information on data protection and privacy, a description of the data sharing functions, and an explanation regarding the incentives. Each letter contained a unique registration code. A reminder mailing was sent after 11 days, including an installation brochure, which walked users through the downloading and registration process step by step. In the second installment, the installation brochure was added to the first mailing. The addition of the installation brochure did not have a significant effect on installation rates. See the online appendix in Kreuter et al. (2018) for full documentation of the invitation materials.

Those willing to participate in the study had three tasks incentivized separately: (1) installing the app from the Google Play store using the QR code, using the link provided in the invitation letter, or by searching for the app name directly in the store, (2) allowing the app to collect sensor and other passive measurements, and (3) answering survey ques-

tions launched through the app at predefined times or triggered by geo-locations. All participants were offered incentives in form of Amazon.de vouchers (see Figure 4.2, bottom) based on the number of points earned (one point = one euro-cent). Vouchers were available for every 500 points earned (5 euro). The total amount a participant could earn varied between 60 and 100 euro, depending on the incentive condition (see section 3.3).

Experimental conditions were assigned randomly to the selected PASS participants. However, as PASS is a household panel, some households received different incentive conditions within the same household. To avoid any confounding due to family members talking to each other, we restricted our analysis to those households that had only one person selected into the IAB-SMART study (N=2,853). This restriction did not negatively affect the distribution of cases to incentive groups (see Figure 4.3, showing roughly equal amounts of cases in each condition within each of the two factors). However, the number of app installations we could use decreased from 687 to 420 app installations.

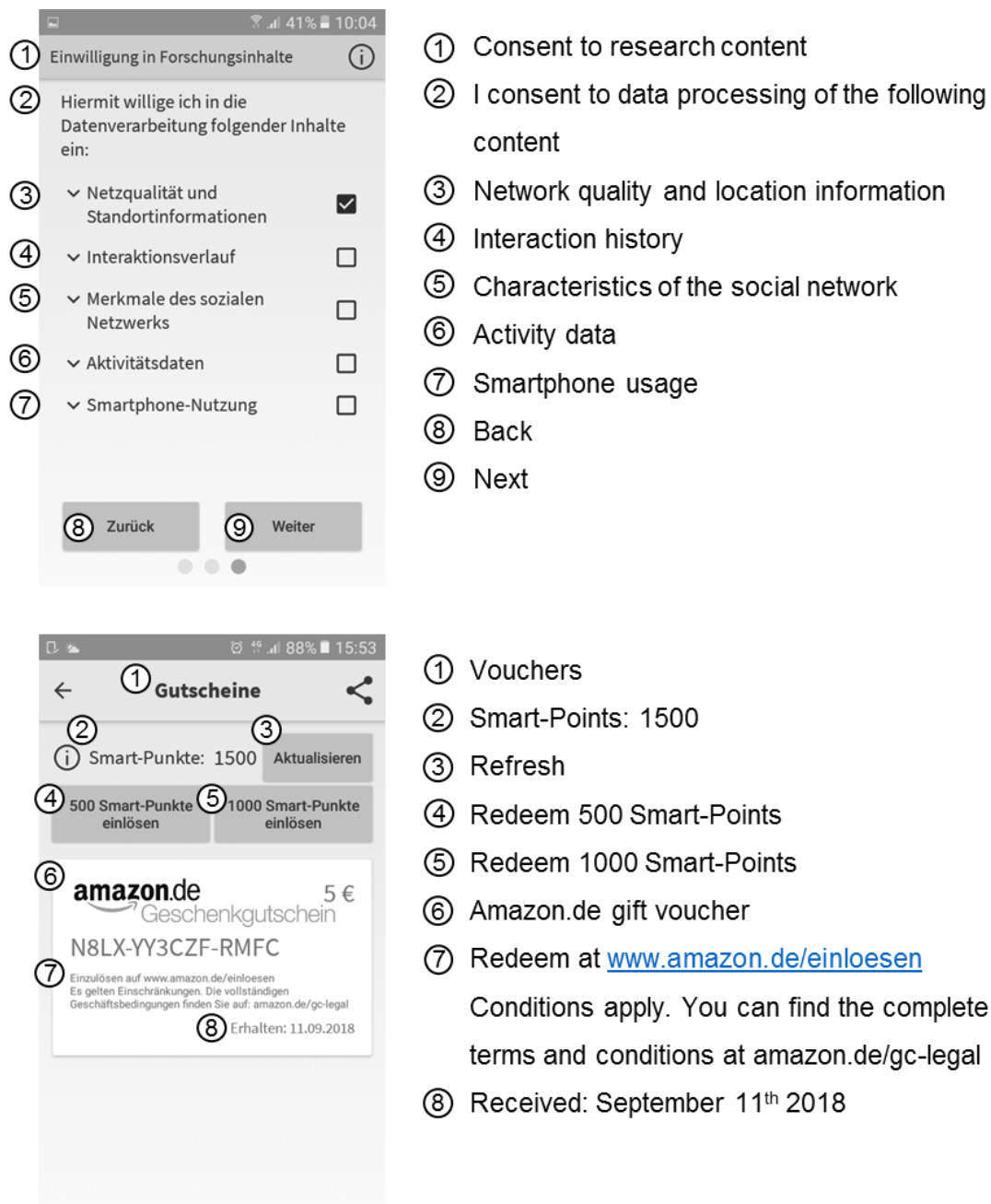


Figure 4.2: Screen-shots showing the five data sharing functions (top) and in-app Amazon.de vouchers (bottom). The app was programmed and offered in German, but direct translations into English are provided in the Figure.

The data passively collected through IAB-SMART are grouped into five data sharing functions. Individuals who installed the app had the opportunity to provide consent to the sharing of their data via any or all of these five functions. Participants could enable and disable data collection in any or all of the five functions at any point in time during the six-month study period (for more details on the consent process, see Kreuter et al. (2018)).

We designed a separate screen for respondents to navigate their data sharing (Figure 4.2 top panel). Allowing Network quality and location information issues a test every half hour where Wi-Fi and mobile network data are collected. Those data allow estimates of the current geo-position of the smartphone. Interaction history records metadata from incoming and outgoing calls and text messages with hashed phone numbers (i.e., taking a string (phone number) of any length and output a nonpersonal random string of a fixed length). The Characteristics of the social network function allowed, if enabled, access to the phone's address book and the classification of contacts into gender and nationalities using the following two websites: genderize.io and www.name-prism.com. Information is pinged to the site, without any names being stored on either providers site. Resulting classification probabilities are retrieved and combined with the hashed phone book contact. The Activity data function collects measurements in two-minute intervals via the smartphone's accelerometer and pedometer. Smartphone usage captures the apps installed on the phone and the start/end-time of each app usage without recording any information about activities done within an app.

4.4.3 Experimental design for incentive study⁶

We conducted a 2x2 experiment on the installation and the function incentives (see Figure 4.3). One random group of participants was promised 10 euro for installing the app, and the other group was promised 20 euro. Independent of the installation incentive (completely crossed), one random group was promised one euro for each function activated for 30 consecutive days, and the other group was promised one euro for each function activated for consecutive 30 days plus five additional euro if all five data sharing functions were activated for 30 days. Consequently, the first group would receive 5 euro and

⁶ The invitation letter contained a flyer, which explained the incentive scheme (for the original flyer and English translation see Appendix Figure 4.13 and Figure 4.14).

the second group 10 euro per month for activating all five data sharing functions. For the sake of simplicity, we refer to the groups of the function experiment as the regular and bonus group. Additionally, all participants receive up to 20 euro for answering survey questions in the app over the field period (10 euro-cent per answered question). Therefore, the maximum promised incentive varies between 60 and 100 euro depending on the assigned group. Participants could redeem their incentives directly in the app as Amazon.de vouchers⁷.

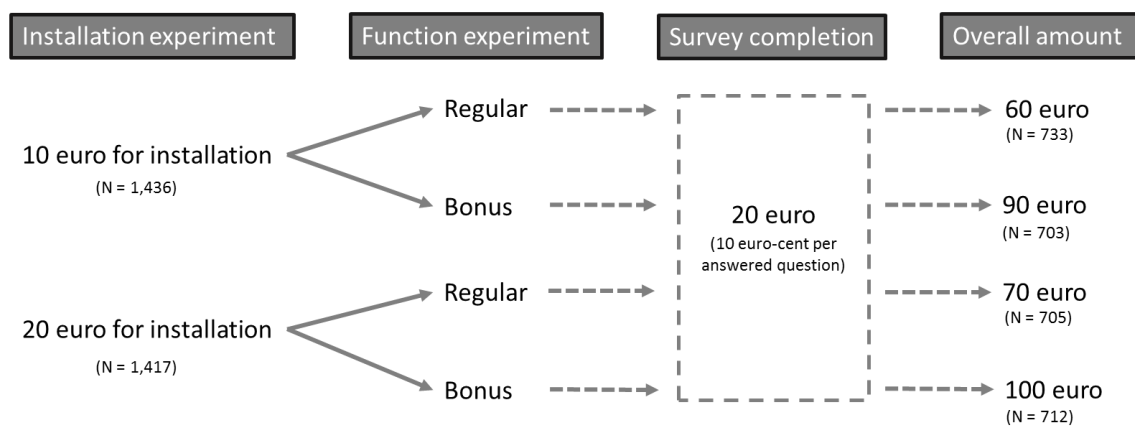


Figure 4.3: Crossed experimental design with maximum incentive amounts for a six-month data collection period (N=2,853)

As the invitation letter mentioned both the individual incentive amounts and the overall maximum amounts, it is conceivable that respondents primarily focused on the overall maximum amounts and less on the differential incentives provided for installation and the app functions. We therefore report both marginal effects for each factor (installation experiment and function experiment, see Figure 4.3) and the combination of the two factors (maximum amounts of 60, 70, 90, and 100 euro).

⁷ When issuing the incentives, we allowed for discretion (e.g., due to network error, etc.). The IAB-SMART app checked whether each data collection function was activated at three random points in time each day. To receive points for activation, the app had to be able to execute the check on at least 10 out of 30 consecutive days. Furthermore, of the days the app was able to execute the check, the function was not allowed to be deactivated on more than three days. However, we did not explicitly mention this point to participants.

4.4.4 Analysis plan

We will analyze the effect of incentives on four outcome variables: (1) installing the app on the smartphone (app installed vs. app not installed), (2) number of initially activated data sharing functions (0–5), (3) deactivating functions during the field period (deactivated a function at least once vs. did not deactivate any functions), and (4) retention (proportion of days out of field period until the app is deinstalled). For each of our four outcome variables, we will proceed as follows. First, we use t-tests, Chi-squared tests, and ANOVAs to examine the main effects of the incentive conditions in our three experiments, i.e., installation experiment, function experiment, and maximum amount, on the outcome variables. Second, we investigate effects of treatment heterogeneity across welfare status⁸ (welfare recipients vs. non-welfare recipients) to evaluate whether vulnerable groups are more affected by incentives. To do so, we examine differences in our outcome variables across welfare recipient status with t-tests, Chi-squared tests, and ANOVAs.

We also analyze to what extent individuals installed the app and cashed-out the incentive without providing any data. We analyze how many points individuals actually redeemed in the experimental groups to study the influence of incentives on costs.

All analyses were conducted using Stata 14.2 (StataCorp. 2015) and R version 3.4.0 (R Core Team 2017). Analysis code and data can be reviewed and accessed on request at the IAB (for more information, see here: <https://www.iab.de/en/daten.aspx>).

⁸ Welfare benefits are paid to all households with insufficient income in which at least one person is of working age (15 to 65) and able to work, regardless of their labor market status.

4.5 Results

4.5.1 App installation

Figure 4.4 shows the overall effects of the different incentive treatments on app installation. A higher installation incentive resulted in a higher installation rate, at 16.4% of those offered 20 euro for installing the app compared to 13.1% of those offered only 10 euro ($\text{Chi}^2=5.8$, $\text{df}=1$, $p=0.01$). We do not observe a marginal effect for the additional function experiment with a 5-euro bonus for activation of all five data sharing functions for 30 consecutive days compared to the regular group ($\text{Chi}^2=0.58$, $\text{df}=1$, $p=0.447$). For the maximum incentive amount, we observe a higher installation rate (bottom panel) for the 70 euro (16.4%) and 100-euro group (16.3%) compared to the 60 euro (12.0%) and 90-euro (14.2%) group. However, the differences are not significant at the 5% level ($\text{Chi}^2=7.53$, $\text{df}=3$, $p=0.057$). As the 70- and 100-euro groups include the 20 euro as an installation incentive, the installation experiment drives the differences in the installation rates by maximum amounts.

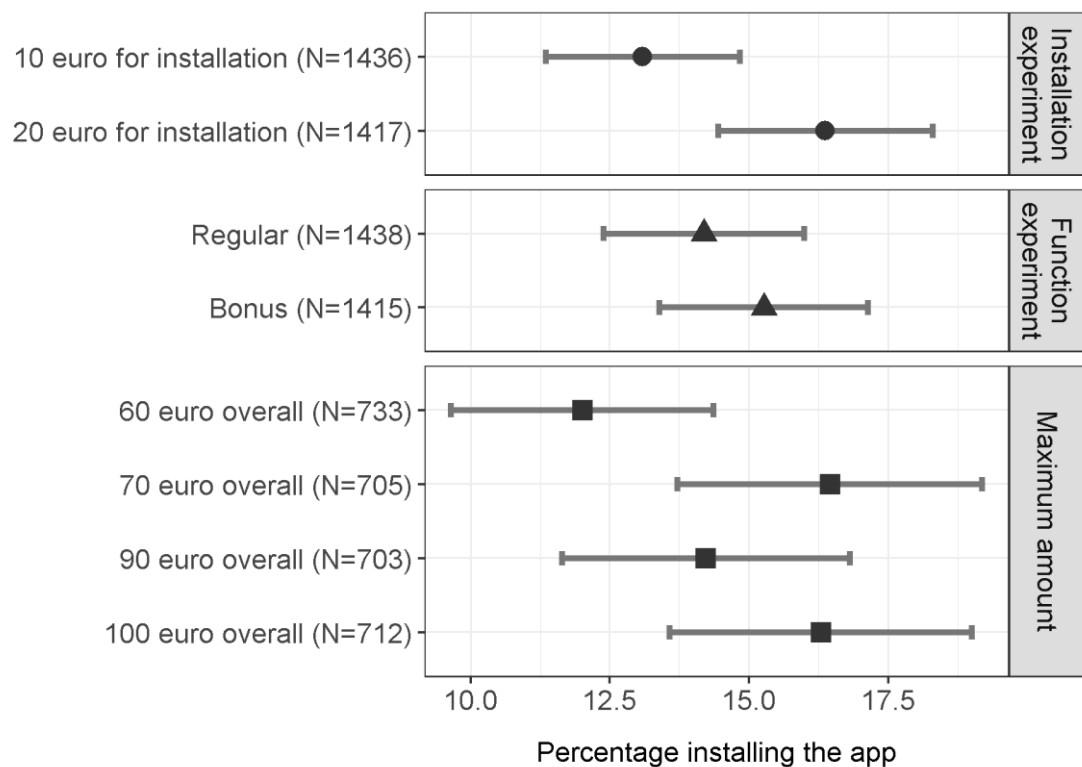


Figure 4.4: Percentage of app installations of invited individuals, with 95% confidence intervals ($N = 2,853$), by incentive condition, maximum amount of incentives and welfare status

We do not find a differential installation rate based on welfare status ($\text{Chi}^2=0.37$, $\text{df}=1$, $p=0.541$). However, to understand a potentially differential effect of incentives on people with different welfare status, we analyze the effects of our incentives for non-welfare and welfare recipients separately (see Figure 4.5). In the top panel of Figure 4.5, we observe that for both welfare status groups, the 20-euro incentive increases the installation rate compared to the 10-euro incentive (3.5 percentage point increase for non-welfare recipients and 2.7 percentage point increase for welfare recipients). We only find, however, a significant effect for non-welfare recipients ($\text{Chi}^2=4.9$, $\text{df}=1$, $p=0.027$), not for welfare recipients ($\text{Chi}^2=0.9$, $\text{df}=1$, $p=0.342$). As the number of cases is lower for the welfare recipient group ($n=729$) than for the non-welfare recipient group ($n=2,118$), the nonsignificant effect for welfare recipients may be tied to smaller sample size. For the function experiment, we observe no difference in the installation rate between the regular and the

bonus function incentive for non-welfare recipients, while for welfare recipients, the bonus incentive leads to a higher installation rate than does the regular incentive. However, this effect is not statistically significant ($\chi^2=3.2$, $df=1$, $p=0.073$), potentially due to the relatively small sample size of welfare recipients. Similarly, there seems to be a linear increase in the installation rate with increasing maximum incentive amount for welfare recipients, but we do not observe a clear pattern for non-welfare recipients. Again, none of the effects is statistically significant (Chi-squared tests; all $p>0.05$).

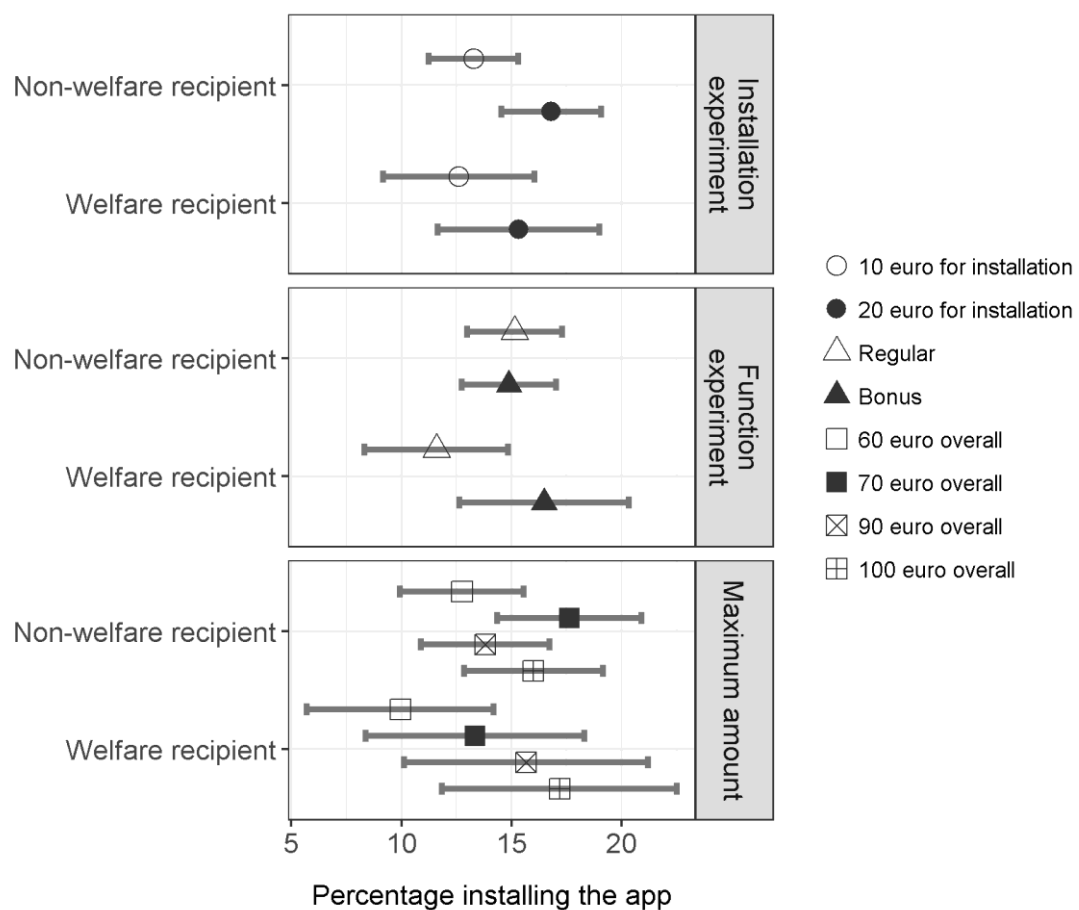


Figure 4.5: Percentage of app installations of invited individuals, with 95% confidence intervals ($N = 2,853$), by experimental groups and maximum amount of incentives, by welfare status

4.5.2 Number of initially activated data sharing functions

When installing the app, participants could choose to activate any of five data sharing functions. One must keep in mind that the general installation of the app is a precondition for being able to receive any incentive for activating a function. In general, we observe high activation rates in all experimental groups (on average, between 4.1 and 4.3 initially activated data sharing functions; see Figure 4.6). The already high activation rate does not change with a higher installation incentive ($t=-0.04$, $df=407.8$, $p=0.971$), the bonus incentive to have all five data sharing functions activated ($t=-0.34$, $df=417.3$, $p=0.735$), or the resulting maximum amounts ($F_{ANOVA}=0.46$; $df=3$; $p=0.707$). The bonus group incentive of the function experiment, however, does not primarily aim to increase the average number of initially activated data sharing functions; instead, it aims to nudge participants to activate all five data sharing functions. Although there is a four-percentage-point difference in activating all data sharing functions between the experimental groups (70.8% for the bonus incentive vs. 66.2% for the regular incentive), this difference is not statistically significant ($\chi^2=0.85$, $df=1$, $p=0.356$).

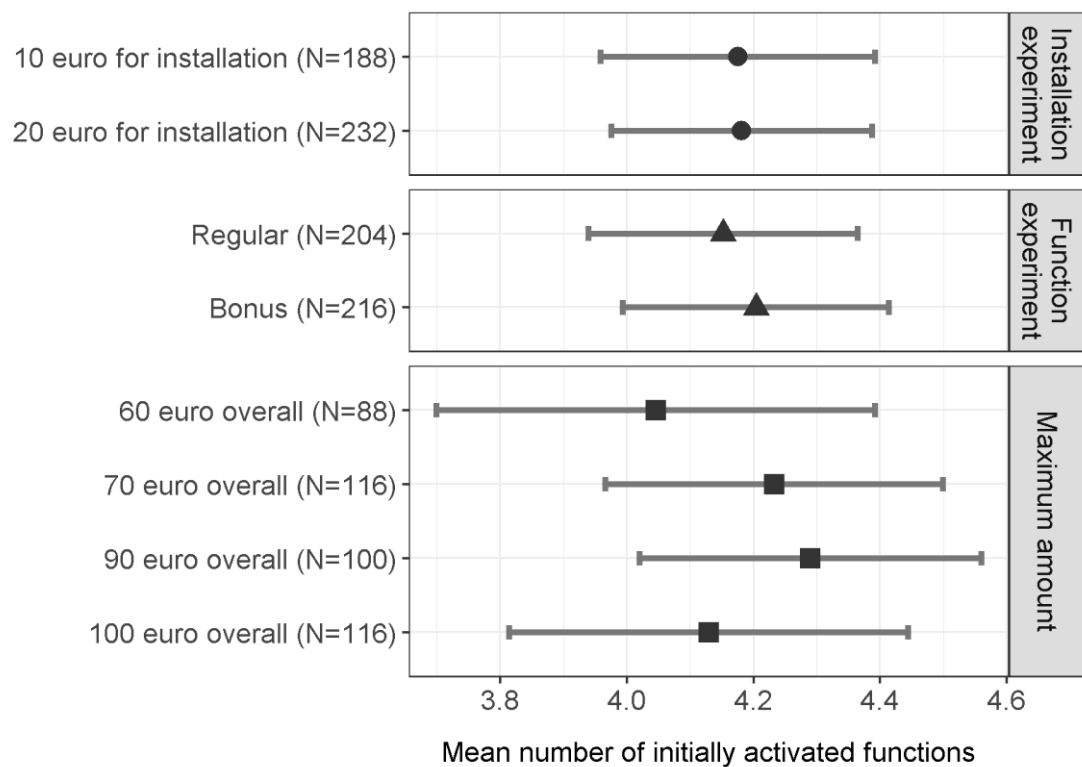


Figure 4.6: Mean number of initially activated data sharing functions, with 95% confidence intervals, conditional on installation, by incentive condition and maximum amount of incentives (N = 420)

We do not find significant differences in the mean number of initially activated data sharing functions by welfare recipient status ($t=-1.58$, $df=145.6$, $p=0.115$). Figure 4.10 (see Appendix) shows the effects of our incentives for non-welfare and welfare recipients. We do not find any significant differences within the welfare subgroups (t-tests and ANOVA; all $p > 0.05$), suggesting that our incentives do not have differential effects within welfare status subgroups.

4.5.3 Deactivating functions

To be compliant with the EU General Data Privacy Regulation (GDPR), we made it easy for participants to change the settings of the data sharing functions in the setting menu of the app during the field period. For the installation experiment and the maximum amount, we expect that higher incentives create a commitment to deactivate fewer data sharing

functions during the field period. For the function experiment, as participants in the bonus group gain an additional five euro by having all five functions activated for 30 consecutive days, the loss associated with deactivating a function is larger than for the regular group. Thus, we expect that fewer participants deactivate a data sharing function in the setting menu for the bonus group.

Only approximately 20% of all participants changed their settings at least once during the field period. Of those participants who changed their function settings at least once, only 31 participants (approximately 7% of 420 participants) deactivated a function at least once. We have very few cases that we can compare over our experimental groups. As a result, we obtain large confidence intervals that overlap and make it hard to find effects (see Figure 4.7). Although there seems to be a pattern of less deactivation with higher installation incentive, bonus incentive (vs. regular incentive), and higher maximum amount, none of these differences were statistically significant (Chi-squared tests; all $p > 0.05$; see Figure 4.7).

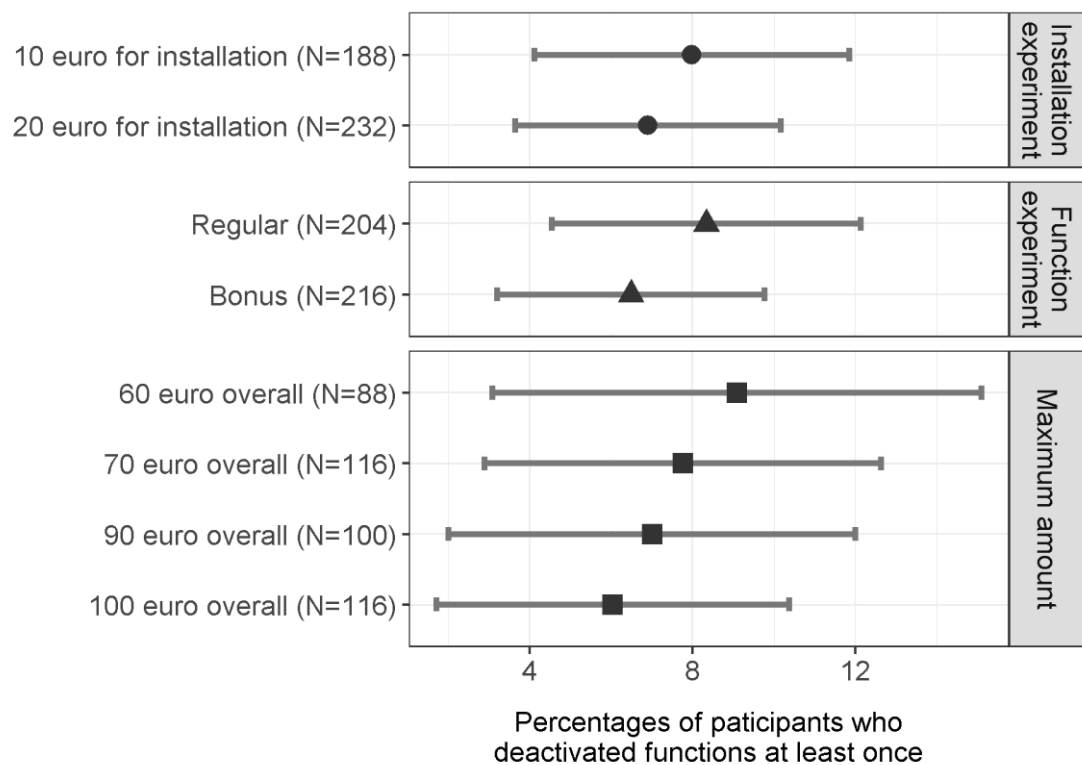


Figure 4.7: Percentages of participants who deactivated their function settings at least once, with 95% confidence intervals, conditional on installation, by incentive conditions, maximum amount of incentives and welfare status (N = 420)

As the proportion of participants who deactivated a function is very low, we do not further compare the effect of incentives within the two welfare status subgroups.

4.5.4 Retention

It was possible for participants to just install the app and provide no or very little information by deinstalling the app rather quickly—but still cashing in an Amazon.de voucher for installing the app. On average, participants kept the app installed for 86% of the field period, meaning that if an individual decided to install the app exactly 100 days before the end of the field period, she kept the app installed for 86 days. Looking at the average percent of days participants stayed in the study (see Figure 4.8), we observe patterns that may suggest that those who received lower incentives deinstalled the app earlier than did those who received higher incentives. Although the effect is not statistically significant

in the installation and function experiments (t-tests; $p > 0.05$), it is for the maximum overall amount ($F_{ANOVA} = 3.13$, $df = 3$, $p = 0.026$). A post hoc test reveals that those receiving up to 60 euro overall stayed on average ten of 100 days fewer than did those receiving 70 or 90 euro (Tukey multiple comparisons of means tests; $p_{adjusted} < 0.05$). However, the difference between the 100-euro maximum amount group and the 60-euro group is not statistically significant (Tukey multiple comparisons of means test; $p > 0.05$).

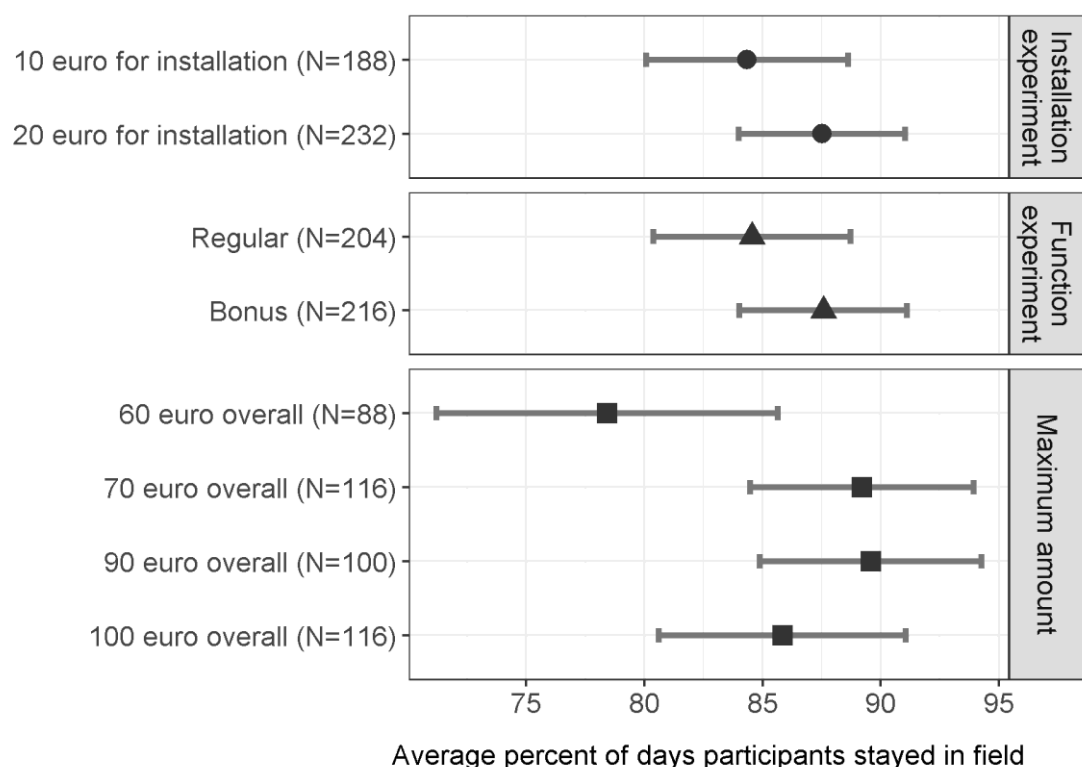


Figure 4.8: Average time participants stayed in field, with 95% confidence intervals, i.e., time between first installation and deinstallation in percent, by incentive conditions and maximum amount of incentives (N = 420)

We do not find a differential retention rate by welfare status ($t = 0.38$, $df = 158.0$, p -value = 0.704). Looking at effects of our incentive experiments within welfare status subgroups (see Figure 4.11 in the Appendix), we find no significant differences (t-tests and ANOVA; all $p > 0.05$) except for the difference between the 60-euro maximum amount (82%) and

the 70-euro maximum amount (91%) for non-welfare recipients (Tukey multiple comparisons of means test; $p_{\text{adjusted}}=0.032$).

Only 20 participants deinstalled the app within one week of installation. Unfortunately, our groups are too small to examine effects of our experimental treatment on the tendency to deinstall the app shortly after installing it.

4.5.5 Analysis of costs

For any study designer, overall costs of data collection are of ultimate interest. In the final section, we therefore analyze how our experimental groups affect costs. To do so, we analyze the average proportion of collected points that participants redeemed.

Overall, 687 individuals installed the app. However, as previously, for the following analyses, we only consider participants from households in which only one member was selected ($N=420$). Of 420 individuals who installed the app, 361 individuals redeemed at least one voucher, and 59 participants did not exchange their points for vouchers. One reason for not redeeming vouchers might be tied to technical issues within the app, i.e., participants received vouchers, but the app failed to store and upload voucher data to the data collection server⁹. Some individuals might not have exchanged any points for vouchers. However, for the sake of simplicity, we assign the value zero to all cases where we have no information on the amount paid out.

⁹ The data for received vouchers and credits has implausible gaps. For example, some participants receive more money in vouchers than they actually collected points. We expect that the missing points appear due to technical errors within the app or communication problems between the app and the server. For our analysis, we assume that missing points are equally distributed over our groups. For our full participant sample ($N=685$), we spent 37,730 euro on vouchers but only have data for received vouchers of 36,070 euro. Therefore, we cannot account for 1,660 euro in voucher. For our restricted participant sample ($N=420$), our data show that we have paid out 21,420 euro in vouchers to our participants. Unfortunately, as we do not know what data is missing for participants who received one invitation per household and for participants who received more than one invitation per household, we have no means of evaluating how large the gap between the redeemed voucher value and the collected points is for our analysis sample. However, it is not possible to receive a higher voucher value than collected points; we therefore conclude that each participant who received more vouchers than collected points actually collected those points without the system storing this information. For example, a participant received a 10-euro voucher, but we have no information on collected points for this participant. To receive a

Individuals may be motivated to participate not only by the incentives we offered but also for other reasons, e.g., to help us collect innovative scientific data or due to curiosity about the new form of data collection. However, the higher the incentive, the higher should be the probability of attracting benefit maximizers whose major motivation to participate is receiving an incentive. If this prediction is true, we should observe that the proportion of redeemed voucher values is lower for smaller incentive groups.

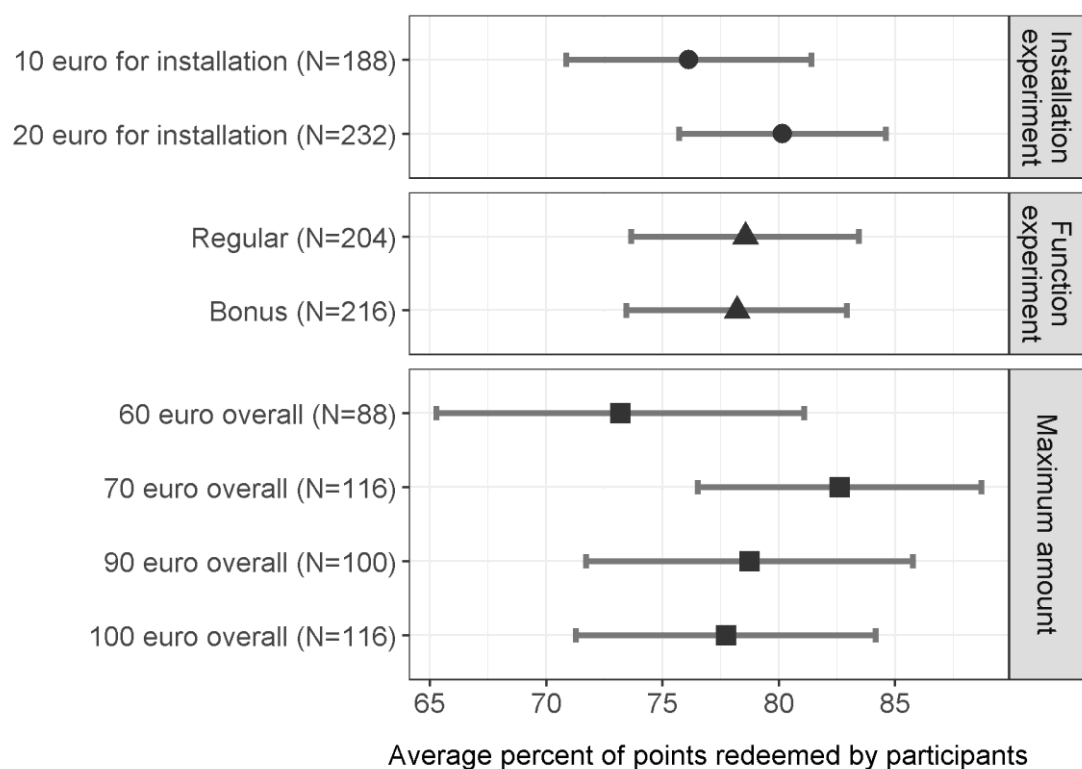


Figure 4.9: Average percent of points redeemed by participants, with 95% confidence intervals, by incentive conditions, maximum amount of incentives and welfare status (N=420)

Overall, participants redeemed 78.8% of their collected points. Figure 4.9 shows the average proportion of points that participants redeemed by our experimental groups. We

voucher, however, (s)he must have collected at least 1000 points; we impute these 1000 points for the participant. Therefore, we adjust our data by adding the missing points to those who redeemed more money in vouchers than collected points. This process reduces the missing points to approximately 26,560 points (265.60 euro).

observe an average proportion of 80.2% of redeemed points for those offered 20 euro for installing the app compared to 76.1% for those offered only 10 euro. However, the difference is not significant ($t=-1.1$, $df=388.5$, $p=0.252$). For the function experiment, we observe a small nonsignificant difference of 0.4 percentage points ($t=0.1$, $df=416.5$, $p=0.917$). We also cannot find any significant effects for the maximum amount ($F_{ANOVA}=1.19$, $df=3$, $p=0.313$).

Our argument that individuals may only participate because we offered incentives may especially affect the group of welfare recipients. Vulnerable groups may be more motivated to participate because of incentives instead of for other reasons such as altruism or curiosity and thus have a higher average proportion of redeemed points compared to non-vulnerable groups who can afford to not redeem all points. However, we cannot find any significant effects of the welfare status on the redeeming of points ($t=0.77$, $df=157.5$, $p=0.443$).

We find no significant difference between our experimental groups within welfare status subgroups (see Figure 4.12 in Appendix; t-tests and ANOVA; all $p > 0.05$), suggesting that incentive conditions do not affect non-welfare and welfare recipients differently with respect to redeeming points.

4.6 Summary

In this article, we investigated the effects of monetary incentives on participation in a study collecting self-reports and sensor data from smartphones based on a completely crossed two-factor experimental design. Target persons sampled from the long-running PASS panel survey in Germany were promised either 10 or 20 euro conditional on installing the app and were promised either one euro for each passive data collection function activated for 30 consecutive days or one euro per function plus a five-euro bonus if

all five data sharing functions were activated for 30 consecutive days. All incentives were provided as Amazon.de vouchers that could be redeemed by participants in the app.

The main finding of this article is that well-known results from the survey literature on the effects of incentives on participation seem to carry over to invitations to share smartphone data. Although the task differs and is less time consuming for participants, we found similar patterns regarding the effects of different amounts of incentives.

A 20-euro installation incentive causes significantly more targeted individuals (16.4 percent) to install an app that passively collects smartphone data than does a 10-euro incentive (13.1 percent). In contrast, paying respondents a five-euro bonus incentive if they grant researchers access to all five data sharing functions neither increases installation rates nor has an effect on the number of data sharing functions activated. This result is surprising, as the potential difference between the regular and the bonus incentive condition over the 180-day field period is three times higher than the installation incentive. We communicated the potential maximum amount to the respondents.

We argued that vulnerable groups such as welfare recipients may be more attracted by the monetary incentives and thus have a higher installation rate. However, for different subgroups defined by welfare status, we find no effect on installation rates.

In the literature, on survey nonresponse, one important issue is whether participants who require more recruitment effort produce data of lower quality, e.g., more item nonresponse. Applied to the research question at hand, one could ask whether participants in the high installation incentive group provide less passively collected data, i.e., initially activate fewer data sharing functions, deactivate functions during the field period, and uninstall the app earlier. For the number of initially activated functions and deactivation of function settings during the field period, we find no evidence that different incentives have an effect. Similarly, for retention, we cannot tie any effect to our installation and

function experiment. What we find is that participants who were offered a maximum incentive of 70 euro and above had the app installed on average for a longer period than did their counterparts in the 60-euro group. Thus, it seems that between the 60- and 70-euro maximum incentive lies a threshold that affects participant's choice to keep the app installed. Our results indicate that the higher incentive does not encourage target persons to collect the incentive and then deinstall the app.

We do not find any effect for welfare status subgroups for the number of initially activated data sharing functions, deactivating function settings during the field period or retention. Furthermore, we do not find evidence of differential effects of incentives across the different subgroups for those three outcome variables. From an ethical perspective, these results are good news because our results suggest that with offering different amounts of incentives, we do not coerce vulnerable groups to share their data. Our analyses, however, may suffer from a low number of cases (especially in the welfare recipient groups) that may mask existing effects. Furthermore, we have to keep in mind that finding no significant differences between vulnerable and non-vulnerable groups does not mean that no individuals felt constrained by being offered an incentive. Offering incentives is ethically problematic, even if one individual was forced to participate because her situation did not allow her to decline the offered incentive.

4.6.1 Limitations and future research

This study comes with several limitations. First, we used data from only one study in Germany, and we do not know yet how results generalize to similar studies with different passive data collection requests or in different countries. All invited smartphone owners had participated in at least one prior wave of a mixed-mode (CAPI and CATI) panel survey. Thus, they are likely to have more-positive attitudes towards scientific studies than does the general population. In addition, a trust relationship between the research institute

conducting the study and the panel respondents has been established, and respondents have become used to receiving a cash incentive of 10 euro per survey wave. These factors might increase willingness to participate in general and modify the reaction to incentives compared to other populations who are not part of an ongoing panel study.

Second, for each type of incentive in our study, there were only two experimental groups and no control group without an incentive. Thus, we have no information on the effect of introducing an incentive versus no incentive or on whether there are diminishing returns when the amount is further increased.

Third, smartphone data collection is still new. All invited PASS respondents are very unlikely to ever have been confronted with a similar request. This novelty factor could lead to different results than might be found once this form of data collection becomes more established. People might become more or less trusting or more or less interested in this kind of research in the future.

Fourth, our study's generalizability may be limited by the type of incentive we used. Throughout the paper, we claim that we provided monetary incentives to participants, but participants never received actual money; instead, they received Amazon.de vouchers. Amazon is the largest online retailer in Germany (Ecommerce News 2017). One could argue that these vouchers are as good as money; however, these vouchers can only be used in the German Amazon online store. They are not usable in brick-and-mortar shops, such as supermarkets, or in other online stores. Individuals may perceive a 10-euro Amazon.de credit as harder to spent than 10 euro in cash. Against this background, the type of incentive may actually modify the effect of the amount of incentive. Furthermore, we have no information on how many of our participants do not have an Amazon.de account and whether those without an account differ from those with an account. Some participants might have preferred different online incentives, e.g., payments through PayPal,

bank transfers, or donations. Therefore, our findings about the effect of incentives are limited to Amazon.de vouchers. Whether different incentives or combinations of incentives are more efficient has to be empirically tested in future research.

Fifth, so far, we have very little knowledge about the value of passively collected smartphone data; nor do participants have a benchmark to estimate their value. Most target persons in our study probably share similar data with commercial providers without pay and without the benefits of anonymization to be able to use certain apps or other online services.

Sixth, our analyses may suffer from small sample sizes. With a higher number of cases in each experimental group, differences between groups may become more clear, and statistical tests may identify effects that are now covered.

Finally, it is not possible to identify whether sample members went to the Google Play Store, looked at the app, but then decided not to download the app. We only have data from individuals who finished the onboarding process, i.e., the process between accessing the Google Play Store and finishing all consent decisions. Only as the onboarding process was finished was the app able to collect data, and we do not have any information about the number of failed onboardings.

References

- Cantor, D., O'Hare, B. and O'Connor, K. (2008). The use of monetary incentives to reduce non-response in random digit dial telephone surveys. In *Advances in telephone survey methodology*, eds. Lepowski, J.M. et al., 471-498. New York: Wiley.
- Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly* 57 (1), 62-79.
- Dillman, D.A., Smyth J.D. and Christian L.M. (2014). *Internet, phone, mail and mixed-mode surveys: The tailored design method*. 4th edition. New York: Wiley.
- Ecommerce News (2017). Top 10 online stores in Germany. Retrieved from <https://ecommercenews.eu/top-10-online-stores-germany/>.
- Felderer, B., Müller, G., Kreuter, F. and Winter, J. (2018): The effect of differential incentives on attrition bias evidence from the PASS Wave 3 incentive experiment. *Field methods*, 30 (1), 56-69.
- Goldenberg, K. L., D. McGrath, and L. Tan. (2009). The effects of incentives on the consumer expenditure interview survey. In *APPOR proceedings*, 5985–99. Hollywood: FL.
- Groves, R. M., M. P. Couper, S. Presser, E. Singer, R. Tourangeau, G. P. Acosta, and L. Nelson (2006). Experiments in Producing Nonresponse Bias. *Public Opinion Quarterly* 70 (5), 720-736.
- Groves, R. M., E. Singer, and A. Corning (2000). Leverage-Saliency Theory of Survey Participation – Description and an Illustration. *Public Opinion Quarterly* 64 (3), 299- 308.

- Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., and Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 11(6), 838-854.
- Jäckle, A., Burton, J., Couper, M.P. and Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Survey Research Methods* 13(1): 23–44. <https://doi.org/10.18148/srm/2019.v1i1.7297>.
- Jacobebbinghaus, P. and S. Seth (2007). The German Integrated Employment Biographies Sample IEBS. In Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Volume 127, pp. 335-342.
- James, J. M., and R. Bolstein. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly* 54 (3): 346–61.
- Keusch, F., Antoun, C., Couper, M. P., Kreuter, F., and Struminskaya, S. (2017). Willingness to participate in passive mobile data collection. *Paper presented at the AAPOR 72nd Annual Conference*, New Orleans, LA.
- Keusch, F. Haas, G.-C., Kreuter, F., Bähr, S., and Trappmann, M. (2018). Coverage error in smartphone data collection. *General Online Research 18*, Cologne, Germany,
- Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., and Trappmann, M. (2018). Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges

- Around Privacy and Informed Consent. *Social Science Computer Review*.
<https://doi.org/10.1177/0894439318816389>.
- Levenstein, Rachel (2010). Nonresponse and Measurement Error in Mixed-Mode Designs. Dissertation. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/78764/rmlev_1.pdf?sequence=1.
- Mack, S., V. Huggins, D. Keathley, and M. Sundukchi (1998). Do Monetary Incentives Improve Response Rates in the Survey of Income And Program Participation? In *Proceedings of the Section on Survey Methodology, American Statistical Association*, pp. 529-534.
- Mercer, A., Caporaso, A., Cantor, D., and Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly* 79(1), 105-129.
- Pfarr, K. (2016). *Incentives*. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. https://doi.org/10.15465/GESIS-SG_EN_001.
- Philipson, T. (1997). Data Markets and the Production of Surveys. *The Review of Economic Studies* 64(1), 47-72.
- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Revilla, M., Couper, M. P., and Ochoa, C. (2018). Willingness of online panelists to perform additional tasks. *methods, data, analyses*. Advance online publication. <https://doi.org/10.12758/mda.2018.01>.

- Sakshaug, J. and Kreuter, F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods* 6(2), 113-122.
- Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse*, pp. 163-177. John Wiley & Sons New York.
- Singer, E., Groves, R. M. and Corning, A. D. (1999). Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly* 63(2), 251-260.
- Singer, E. and Kulka, R. A. (2002). Paying respondents for survey participation. In M. V. Ploeg, R. A. Moffitt, and C. F. Citro (Eds.), *Studies of welfare populations: Data collection and research issues*. 105–28. Washington, DC: National Academy Press.
- Singer, E., van Hoewyk, J., Gebler, N., Raghunathan, T. and McGonagle, K. (1999). The Effect of incentives on Response Rates in Interviewer-Mediated Surveys. *Journal of Official Statistics* 15(2), 217-230.
- Singer, E. and Ye, C. (2013). The Use and Effects of Incentives in Surveys. *The Annals of the American Academy of Political and Social Science* 645(1), 112-141.
- StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.
- Sugie, N. F. 2018. Work as Foraging: A Smartphone Study of Job Search and Employment after Prison. *American Journal of Sociology*. 123(5): 1453-1491.

- Toepoel, V. (2012). Effects of Incentives in Surveys. In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences*, pp. 209-223.
- Wenz, A., Jäckle, A., and Couper, M.P. (2017). Willingness to use mobile technologies for data collection in a probability household panel (Understanding Society Working Paper 2017-10). Available at: <https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2017-10.pdf>.
- Willimack, D. K., Schuman, H., Pennell, B.-E. and Lepkowski, J. M. (1995). Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to- Face Survey. *Public Opinion Quarterly* 59(1), 78-92.

Appendix

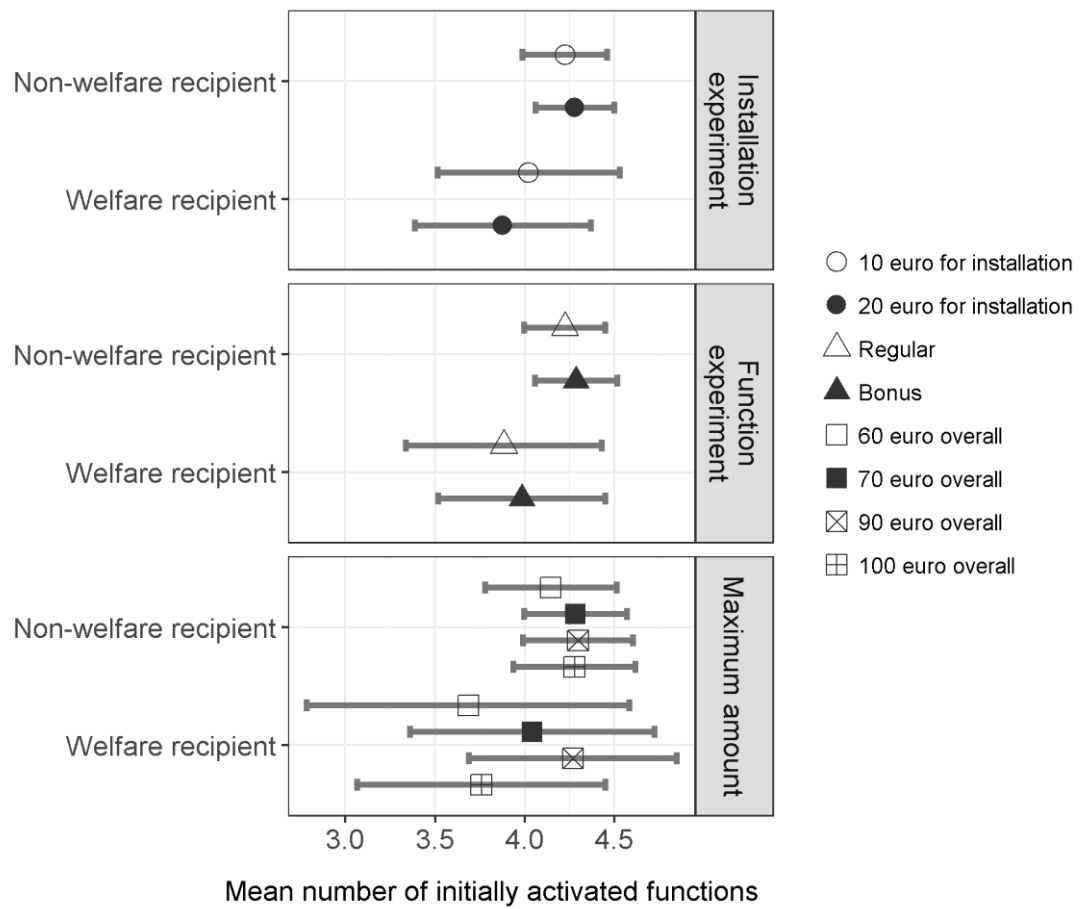


Figure 4.10: Mean number of initially activated data-sharing functions with 95% confidence intervals, by incentive conditions and maximum amount of incentive, by welfare status (N = 420)

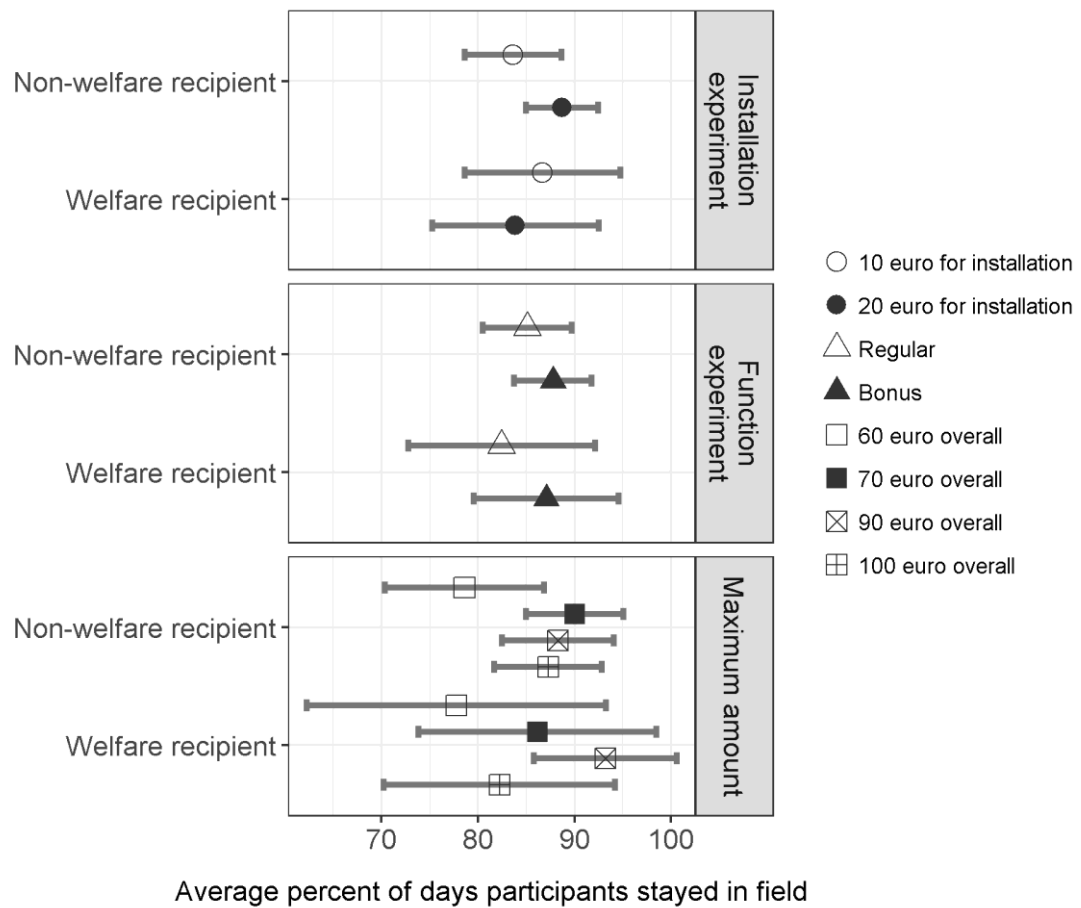


Figure 4.11: Average time participants stayed in field with 95% confidence intervals, i.e., time between first installation and deinstallation in percent, by incentive conditions and maximum amount of incentive, by welfare status (N = 420)

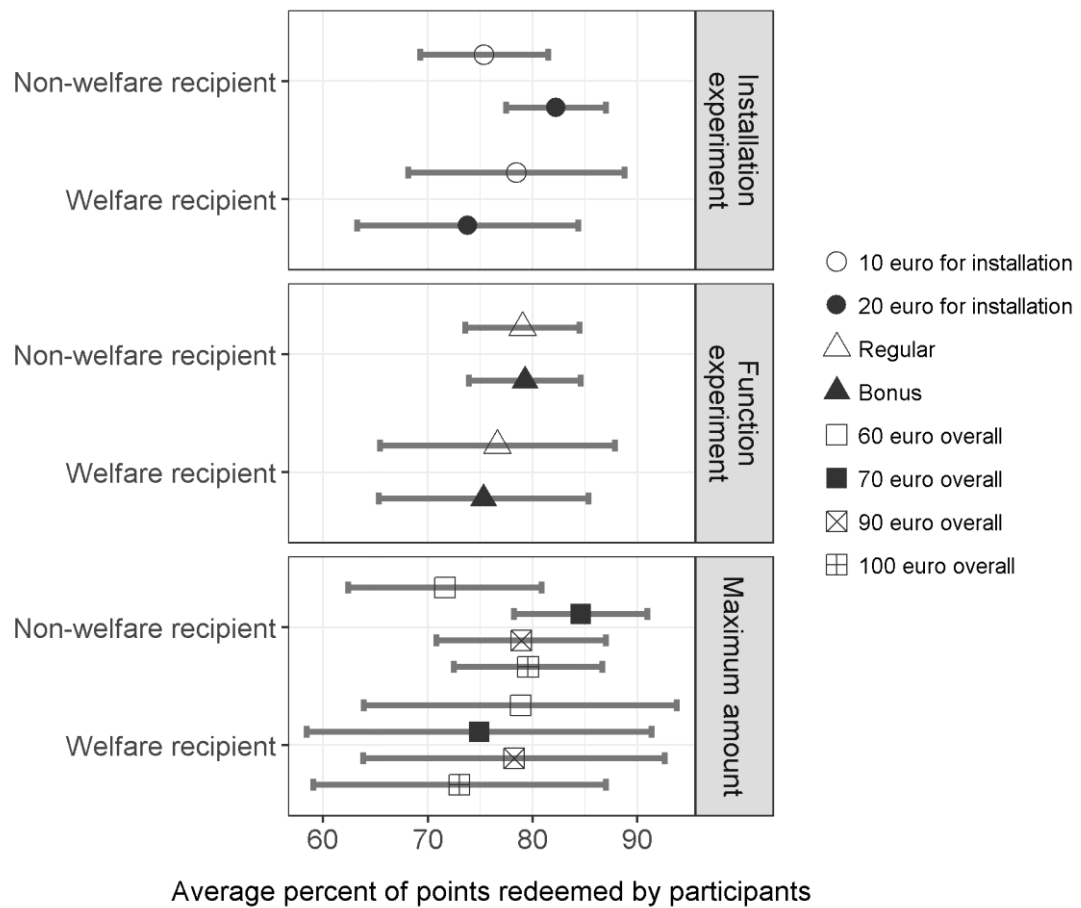


Figure 4.12: Average percent of points redeemed by participants with 95% confidence intervals by incentive conditions and maximum amount of incentive, by welfare status (N=420)



**Der Forschung helfen und
Amazon.de Gutscheine* erhalten!**



App downloaden

Als erstes Dankeschön für die Teilnahme an unserer Studie erhalten Sie von uns einen **10/20 Euro** Amazon.de Gutschein nach erfolgreichem Download der IAB-SMART-App. Diesen können Sie sich in der App unter dem Menüpunkt „Gutscheine“ sofort ausgeben lassen.

Innerhalb der App können Sie Punkte sammeln**. Haben Sie 500 Punkte gesammelt, können Sie diese in einen Fünf Euro Amazon.de Gutschein umwandeln. Punkte sammeln Sie durch

- Aktivierung von Funktionen der App
- Beantwortung von Fragen

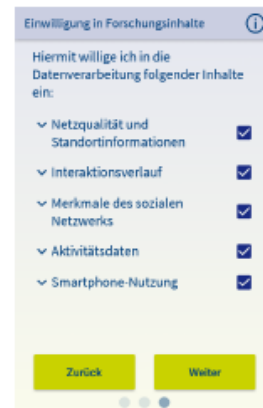
Insgesamt können Sie so weitere bis zu **60/70/90/100 Euro** in Gutscheinen erhalten.

Funktionen aktivieren

Bei der Installation der App, werden Sie gefragt, welche Funktionen der App Sie aktivieren möchten. Für jede aktivierte Funktion erhalten Sie alle 30 Tage 100 Punkte. Später können Sie jederzeit die einzelnen Funktionen unter dem Menüpunkt „Einstellungen“ aktivieren bzw. deaktivieren.

Beispiel: Haben Sie die Funktion Aktivitätsdaten und Interaktionsverlauf aktiviert erhalten Sie alle 30 Tage 200 Punkte

Haben Sie alle Funktionen gleichzeitig aktiviert, erhalten Sie alle 30 Tage 500 Punkte. \ Haben Sie alle Funktionen gleichzeitig aktiviert, erhalten Sie alle 30 Tage 500 Extrapunkte, also insgesamt 1000 Punkte.



Bitte beachten Sie, dass bei der Deaktivierung einer Funktion Ihre angesammelten Tage verfallen.

Alle Funktionen, wenn sie aktiviert sind

Fragen beantworten

Im Zeitraum der gesamten Studie haben Sie die Chance Fragen zu arbeitsmarktrelevanten Themen zu beantworten. Für jede Frage, die Sie beantworten, erhalten Sie Zehn Punkte.

Punkte einlösen

Ihren aktuellen Punktestand finden Sie in der App unter „Gutscheine“. Sobald Sie 500 Punkte gesammelt haben, können Sie diese in einen Fünf Euro Amazon.de Gutschein umwandeln.

*Es gelten Einschränkungen. Die vollständigen Geschäftsbedingungen finden Sie auf: amazon.de/gc-legal.

**diese Aktion gilt bis zum 31.07.2018

Figure 4.13: Original (German) voucher flyer; experimental conditions are marked in red.



Support research and receive
Amazon.de vouchers*!



Download the App

The first gesture of appreciation for participating in our study is a **10/20 Euro** voucher from Amazon.de that you will receive after successfully downloading the IAB-SMART-App. You will immediately be able to access the voucher selecting menu item „Vouchers“.

The app comes with a points-based system**. For each 500 points you get a five euro Amazon.de voucher in exchange. You can earn points by

- Activating extended features in the app
- Answer questions

Overall, you can collect vouchers in the total amount of up to **60/70/90/100 Euro**

Activating extended features

During the installation process, the app asks you to activate extended features. For each feature you activate, you will receive 100 points after each 30 days. You are free to activate or deactivate the features to a later date by selecting the menu item „Settings“.

For Example: If you activate the feature „Activity Data and Interaction History“ you will gain 200 points every 30 days.

If you activate all five features, every 30 days you will receive 500 points./ If you activate all five features, every 30 days you will receive 500 extra points, that means overall 1000 points!

Please note that the gathered days you achieved will decay when you disable this feature.



All features when they are activated.

Answer questions

While the study takes place, you will have the opportunity to complete surveys to labour market-related issues. For each answered question, you will receive ten points.

Redeem points

To check your actual status points select „Vouchers“ in the app menu. As soon as you reach 500 points you can convert them into a five euro voucher from Amazon.de.

* Limitations apply. For full details and terms of conditions please visit: amazon.de/gc-legal.

**This offer does apply until 31.07.2018

Figure 4.14: Voucher flyer (English translation); experimental conditions are marked in red.

5 Using Geofences to Collect Survey Data: Lessons Learned From the IAB-SMART Study

5.1 Abstract

Within the survey context, a geofence can be defined as a geographical area that triggers a survey invitation when an individual enters the area, dwells in the area for a defined amount of time or exits the area. Geofences may be used to administer context-specific surveys, such as an evaluation survey of a shopping experience at a specific retail location. While geofencing is already used in other contexts (e.g., marketing and retail), this technology seems so far to be underutilized in survey research. We implemented a geofence survey in a smartphone data collection project and geofenced 410 job centers with the Google Geofence API. Overall, the app sent 230 geofence-triggered survey invitations to 107 participants and received 224 responses from 104 participants. This article provides an overview of our geofence survey, including our experiences analyzing the data. We highlight the limitations in our design and examine how those shortcomings affect the number of falsely triggered surveys. Subsequently, we formulate the lessons learned that will help researchers improve their own geofence studies.

5.2 Introduction

Designed for the purpose of survey data collection, a geofence can be defined as a geographical area that triggers a survey invitation when an individual enters the area, dwells in the area for a defined amount of time or exits the area. For a geofence to work, the individual needs to carry a device, such as a smartphone, that collects geolocation data and allows geofence software to run within an app that triggers the survey invitation. The geofence software can identify if the individual is inside or outside the geofence.

In market research, geofences are used to collect real-time feedback about a store or other establishment aimed at reducing the recall bias of costumers (Greenwood 2017).

Geofences might also be used in combination with ecological momentary assessments (EMAs, e.g., Stone and Shiffman 1994). Usually, EMAs consist of asking participants about their current affect or behavior at random points in time during a day. Geofences allow researchers to target these questions about people's moods or behaviors when they are at a specific location (e.g., at school, the work place, a fitness studio). For instance, Wray et al. (2019) used smartphone geofences to evaluate if specific characteristics of locations such as bars are associated with consuming more alcohol when visited by study participants.

However, to date, the literature on geofences in surveys is very sparse, and no clear study design guidelines exist that would help researchers to avoid certain pitfalls when employing this technology for survey research.

We conducted a feasibility study where we geofenced 410 job centers in Germany to assess whether geofence surveys can provide researchers with insightful data on formal job search methods. Usually, the Panel Study "Labour Market and Social Security" (PASS) collects data on formal job search methods with a yearly telephone or face-to-face survey in Germany (Trappmann et al. 2019). One dimension of the formal job search methods that PASS assesses is welfare recipients visiting a job center. Asking respondents about their experience during job center visits once a year may bias estimates if respondents visit the job center multiple times during that year. During the PASS interview, respondents have to summarize their experience over all visits in a given year. Furthermore, since job center visits may have happened almost a year ago at the time of the interview, responses are likely to suffer from recall bias (see Tourangeau et al. 2000). Geofences offer the possibility to collect information on respondents' feelings directly after each job center visit, that means, for each job center visit, we get a timely estimate of the current visit.

During analysis of the data from our feasibility study, we noticed several challenges that are easy to overlook when designing a geofence survey. Since the literature provides little guidance on how to conduct geofence-triggered survey data collection, this article serves as a summary of different challenges that survey researchers should consider when conducting a geofence survey. The geofence study is part of a larger app data collection project (Kreuter et al. 2018). We first provide an overview of the main study and describe our geofence survey design. We then report the number of triggered surveys and responses, followed by an evaluation of our geofence study. Finally, we list the lessons learned from our study that will help future geofence studies to improve their designs.

5.3 Design

The IAB-SMART study uses a smartphone app to collect data for labor market research from the smartphones of participants. The app was designed to collect passive smartphone data and to deliver short surveys. In January 2018, we invited 4,293 participants of the Panel Study “Labour Market and Social Security” (PASS) via a postal letter to install the IAB-SMART app on their smartphones, respond to survey questions and passively share data over a period of six months.

PASS is a household panel survey based on a probability sample of the residential population aged 15 and above in Germany with annual waves of data collection (Trappmann et al. 2019). The goal of PASS is to facilitate research on unemployment, poverty and the receipt of state transfers. The questionnaire focuses, among other topics, on income sources, deprivation, (un)employment, job search behavior, social inclusion and attitudes towards the labor market. A dual sampling frame (population registers and welfare benefit recipient registers) is used in order to oversample welfare benefit recipients (for more information, see Trappmann et al. 2013). The data collection mode of PASS is a sequential mixed-mode combination of computer-aided personal and telephone interviews.

Overall, 13,703 respondents participated in wave 11 in 2017. Invitation to the IAB-SMART study was restricted to respondents aged 18—64 ($n=11,208$) who had reported owning an Android-operated smartphone ($n=6,544$), conducted their wave 11 PASS interview in the German language ($n=5,826$) and agreed to be re-contacted for the panel ($n=5,771$). We only invited Android smartphone users because extensive passive data collection is restricted under iOS (the operating system of Apple iPhones). The shares of other operating systems are too small to justify the effort to program additional apps. Keusch et al. (2020a) evaluated how smartphone owners as well as android and iOS smartphone owners differ from the general population in Germany. The authors find that the likelihood of owning an Android smartphone increases with being male, younger and with a higher formal education level. Out of the 5,771 eligible respondents, 4,293 were randomly selected and invited to participate in the IAB-SMART study with a postal letter and one reminder. Overall, 685 of the invited PASS participants installed the app (Keusch et al. 2020b).

During the installation process, individuals could decide if they wanted to allow the IAB-SMART app to passively collect data by enabling up to five data collection functions: (1) network quality and location information, (2) interaction history, (3) social network characteristics, (4) activity data, and (5) smartphone usage. Withdrawing consent was possible at any time in the app's setting menu. For the purpose of this specific study, we only used information from the first function, and we only did this to verify information about geofences (see below). More details on the other functions, including consent rates, can be found in Kreuter et al. (2018).

“Network quality and location information” app function

If an individual decided to enable the “Network quality and location information” function, they allowed the app to collect the location of the smartphone every 30 minutes and

to trigger surveys via the geofences. Note, however, that the geofencing did actually happen outside this custom-made function (see below). However, to receive a geofenced survey in the app, the function needed to be enabled by the participant. Out of the 680 participants of the IAB-SMART study, 577 participants (87.4 %) successfully shared at least one geoposition with us during the data collection period and 209 participants (30.6 %) shared at least one geoposition per day for over 180 days. To collect geopositions, four different methods were used, with each method acquiring data with different accuracy: (1) GPS (median accuracy: 12 meters), (2) mobile carrier network (median accuracy: 20 meters), (3) WiFi (median accuracy: 30 meters), and (4) cell tower database (median accuracy: 930 meters). With each 30-minute measurement, the app tried to collect the most accurate geoposition available (see Bähr et al. 2020). We used this information to verify which geofence triggered a survey.

Job center geofences

We specified 410 geofences distributed across Germany for our study. Each geofence was defined as an area with a 200-meter radius around a job center. Job centers are agencies responsible for the provision of welfare benefits for people aged 15–64 who are able to work. In Germany, this welfare benefit is called Unemployment Benefit II and is available to all households with an insufficient income, irrespective of the labor market status of their members, as long as at least one member is aged 15–64 and able to work. Job centers administer the payments but also have the task to support recipients in finding employment, providing them with job offers, and offering training or active labor market policy programs. For long-term unemployed who have lower chances to reenter the labor market, such training and programs can focus on stabilizing their life situations and improving their employability. Welfare benefit recipients usually visit their local job centers at regular intervals. These visits can happen for two different purposes: (1) visits to file

and discuss claims (administrative meetings) or (2) visits to improve labor market and life situations (consulting meetings).

To prevent falsely triggered surveys (e.g., due to passing by a center), we defined a minimum duration of 25 minutes within the geofence before a survey was triggered. Those 25 minutes are based on a plausible guess on the minimum length of stay at a job center. However, we may have missed some job center visits that took less than 25 minutes. The app would administer a short survey upon exiting the geofence, asking the participant if she had a consulting meeting, and, if so, their experience with the meeting. The survey was triggered after a participant exits the geofence to prevent participants from responding to the survey during their job center visits.

To identify when a participant stayed for at least 25 minutes in a geofence and then exited it, we used the Google Geofence API. The Google Geofence API measures three events (see Figure 5.1): (1) whether an individual enters, (2) dwells and (3) exits the geofence. In our use case, the Google Geofence API only documented how long an individual dwelled in the geofence and when she exited. For the sake of simplicity, we will use the term *visited* to describe the procedure of dwelling for at least 25 minutes in a geofence and exiting it.

Note that the Google Geofence API operated independently of our custom “Network quality and location information” function in the IAB-SMART app, and it used Google Services for geopositioning to identify whether an individual visited a geofence. To preserve participants’ privacy, we did not save any geolocation data measured from the Google Geofence API or any other apps. This implies that we did not collect data on which specific geofence triggered a survey. However, to evaluate how well the geofences worked in terms of identifying if a participant was within a geofence, we can use the timestamp of the survey trigger and the geolocation information from our custom 30-

minute interval as an approximation of whether a participant was or was not within a geofence at the time the survey invitation was sent (see the Results section).

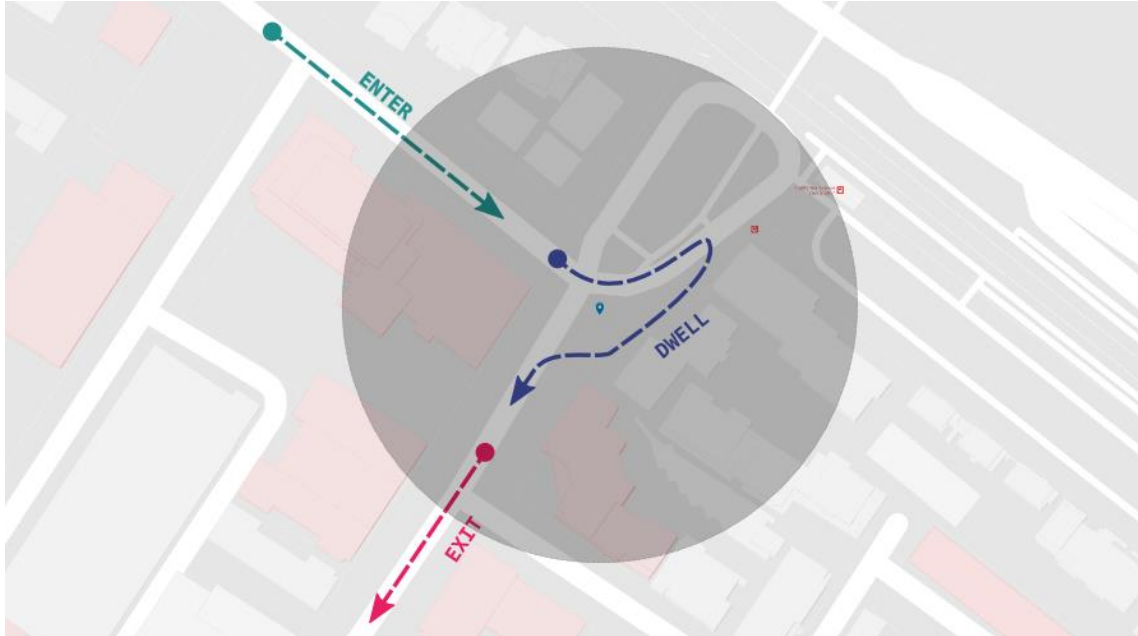
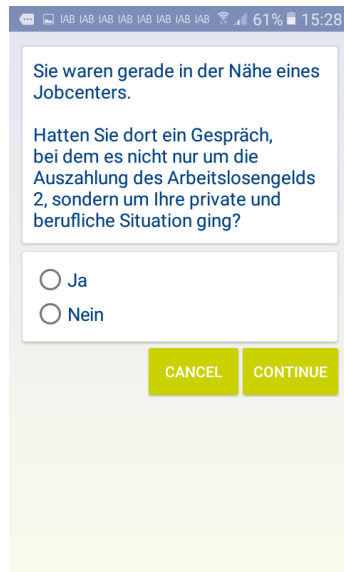


Figure 5.1: The three events (Enter, Dwell and Exit) that the Google Geofence API measures (source: <https://developers.google.com/location-context/geofencing/>, accessed: January 12, 2020).

The Google Geofence API has a limit of 100 geofences per user and device. To circumvent this limitation, each user was dynamically assigned to 100 geofences depending on their current spatial area.

Survey data

Upon visiting a job center geofence, the Google API triggered a survey invitation through the IAB-SMART app asking about the job center visit. The first question was used to verify that the participant had a consulting meeting (see Figure 5.2).



Sie waren gerade in der Nähe eines Jobcenters.

Hatten Sie dort ein Gespräch, bei dem es nicht nur um die Auszahlung des Arbeitslosengelds 2, sondern um Ihre private und berufliche Situation ging?

☐ Ja

☐ Nein

CANCEL CONTINUE

You were just near a job center.

Did you have a conversation there that was not only about the payment of unemployment benefit 2, but about your private and professional situation?

- ☐ Yes
- ☐ No

Figure 5.2: Verification question that appeared as the first question upon accessing the geofence survey with the translation on the left.

If a participant answered the question with no, no follow-up questions were asked; if a participant answered the question with yes, ten follow-up questions were asked evaluating the consulting meeting with the placement officer (see Appendix Figure 5.6 for the full wording). In the IAB-SMART app, participants were incentivized to allow passive data collection and respond to the short surveys (the job center survey was one of a total of twelve different survey modules programmed into the app). For each answered survey question, participants received an incentive of 10 points, that is, participants received 10 points by answering the first question with “No”, and 110 points by answering the first question with “Yes” and completing the entire survey module. Once participants reached 500 points, they could convert the points to amazon.de vouchers; 500 points equaled a 5 Euro voucher (for more information about the incentives, see Haas et al. 2020).

5.4 Results

To assess how well the geofence study worked, we organize the presentation of our results in two sections. First, we present the number of triggered invitations and responses for all

job center geofences as quantitative measures of how often the geofences triggered a survey in the IAB-SMART study. Second, we discuss the challenges by qualitatively evaluating (1) how considering the operation times of job centers would have affected the number of triggered surveys and responses, (2) if the participant visited a valid geofence, (3) on which day the survey was answered and (4) how well the geofence trigger worked by assessing if the location of our custom function measurement was within a geofence shortly before the time that a survey was triggered. We use data from the custom “Network quality and location information” function and the responses to the in-app survey questions for this purpose.

5.4.1 Number of triggered surveys and responses

If a participant visits a geofence, the IAB-SMART app triggers a survey invitation. Overall, the IAB-SMART app sent 230 geofence-triggered survey invitations to 107 participants. Table 5.1 shows that the majority of participants (62) received only one, 18 participants received two and 26 participants received more than two geo-triggered survey invitations, including one participant who received nine invitations. Overall, 104 out of the 107 (97.2%) IAB-SMART participants who received a geofence-triggered survey invitation responded at least once. In terms of invitations, 224 out of a total of 230 (97.4%) survey invitations that were sent led to a response by a participant. Out of these, participants reported 56 times (25.0%) that they had a consulting meeting in the job center.

Table 5.1: Number of IAB-SMART participants by the number of triggered surveys

Number of triggered surveys per IAB-SMART participant	1	2	3	4	5	6	7	8	9
Number of IAB-SMART participants (N=107)	62	18	9	5	2	5	4	1	1
In percent	57.9	16.8	8.4	4.7	1.9	4.7	3.7	0.9	0.9
Sum of triggered surveys (N=230)	62	36	27	20	10	30	28	8	9
In percent	27.0	15.7	11.7	8.7	4.3	13.0	12.2	3.5	3.9

Figure 5.3 shows the positions of all geocoded job centers in the IAB-SMART app. The size of the marker indicates the number of triggered surveys per job center. Overall, 79 of the 410 (19.3%) implemented job center geofences were triggered at least once, whereas the number of triggered invitations per job center ranges from one to 15 times (see Table 5.2).

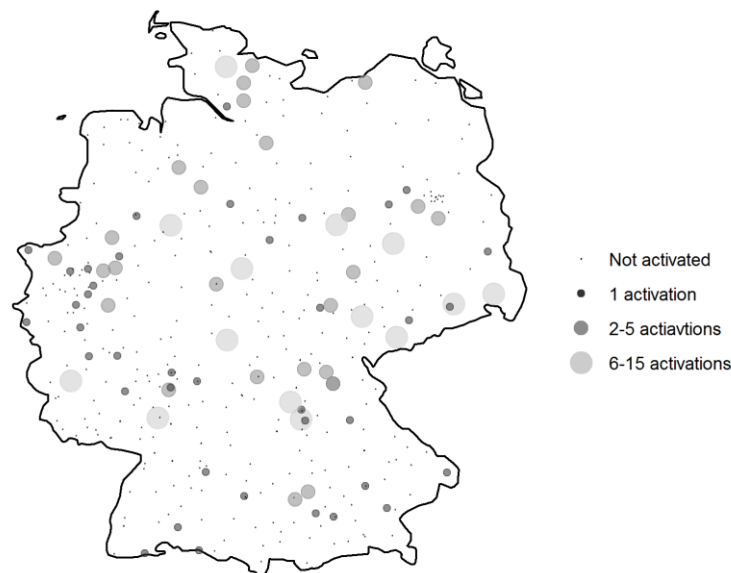


Figure 5.3: Implemented job center geofences in Germany (N=410)

Table 5.2: Number of triggered surveys by the number of job centers

Number of triggered surveys per job center	1	2	3	4	5	6	7	8	9	15
Number of job centers (N=79)	39	15	4	4	3	5	1	3	3	2
In percent	49.4	19.0	5.1	5.1	3.8	6.3	1.3	3.8	3.8	2.5
Sum of triggered surveys per job center (N=230)	39	30	12	16	15	30	7	24	27	30
In percent	17.0	13.0	5.2	7.0	6.5	13.0	3.0	10.4	11.7	13.0

Only after the end of our data collection did we notice that five pairs of job centers were very close to each other and thus had overlapping geofences. The closest distance between two job centers in the IAB-SMART app is 167 meters. If a participant visits the overlapping space of two geofences, it is not clear which geofence triggered a survey and, thus, it is not clear which triggered surveys belong to which job center. We did, however, not find a geolocation in our data that indicates that any participant actually dwelled within the overlapping area of two geofences.

5.4.2 Challenges

Operation times of job centers

Our design did not consider the operation times of job centers, which made geofence triggered surveys possible at times when job centers were closed. For the sake of simplicity, we assume that job centers operate from 7 am to 7 pm on business days. The actual opening times may vary slightly between job centers, but these were the maximum opening hours found in a small data collection from the websites of a sample of 15 job centers. Overall, we find 45 (19.6%) triggered surveys on the weekend. Additionally, we find 14 (6.0%) triggered surveys before 7 am or after 7 pm on weekdays. As a result, we have 59 (25.6%) clearly false triggered surveys that could have been avoided by considering the

operation times of job centers. This can even be considered a conservative estimate since we chose the maximum opening times.

Each participant who received a survey invitation at the time the job center was closed and responded to the survey (N=59) should have answered “No” to the first survey question asking if a consulting meeting took place. However, we find that in eight surveys (13.6%), respondents reported that they had a consulting meeting in the job center, which are probably false reports. If we sum up the incentive costs of those false triggered surveys, we obtain an amount of 13.9 Euros, which is approximately 18% of the overall incentive costs for the geofence surveys (77.4 Euros). While the monetary consequences are negligible in our study, geofence studies with larger sample sizes may benefit from the cost savings through considering the opening times. Additionally, this points to a potential measurement error problem induced by the incentive structure. Identified false reporters can be interpreted as a reversal of an effect termed ‘motivated underreporting’, where respondents answer filter questions in such a way that they avoid lengthy follow-ups (Eckman et al. 2014). When confronted with the option to earn extra money per question, respondents might tend to choose longer paths through the survey.

Valid geofences per participant

In our design, each participant was able to access each geofence and trigger a survey. In practice, however, each participant is assigned to one job center based on the participant’s home address located within the administrative area of the job center. We do not know the administrative areas of our job centers. For the sake of simplicity, we thus assume that the responsible job center is within a radius of less than 100 kilometers of the participant’s home address. If the distance between a job center geofence that triggered a survey and the home address is greater than 100 kilometers, we can assume that this person walked randomly into the geofence (e.g., during a business or leisure trip).

To infer home addresses, we use our collected geolocation data from the custom “Network quality and location information” function. First, we assume that most individuals stay more nights at home than anywhere else (i.e., even if individuals work night shifts, they should be more at home than at other places). Second, we round the coordinates of the location measurement to the 3rd digit after the decimal point and identify the rounded location that appears the most often from 8 pm to 6 am over all days of data in the study. Third, we calculate the average of the unrounded location measurements to obtain an approximate home address of the participant.

Overall, we find nine triggered invitations that are more than 100 kilometers away from the participants’ home addresses and are likely implausible. We find that in one of these nine triggered surveys, respondents stated that they had a consulting meeting.

Availability of the survey invitation

With a few exceptions, all surveys sent through the IAB-SMART app were available to participants for seven days after the initial invitation. All geofence survey questions, however, contained the word “TODAY” to reference the day of the geofence visit. This might compromise the validity of the survey responses for participants who did not respond on the day the invitation was sent. We rely on their implicit understanding that the questions refer to the date of the job center visit that triggered the survey.

Comparing the timestamp of the survey invitation with the timestamp of the survey response, we find that for 74 of 224 responded surveys (33.0%), the day of the survey invitation does not match the day of the survey response.

Evaluating the geofence survey trigger

A geofence should only trigger a survey when an individual visits that geofence. In practice, however, the geofence may malfunction in two ways. First, a survey might be triggered even though the geofence was not visited (false positive). We are concerned about this kind of error because each additional survey invitation may increase the respondent's burden to participate (Bradburn 1978). Furthermore, the content of the survey may be out of context and thus increase the burden. In addition, when individuals receive an incentive for responding to a geofence survey, each survey invitation increases the data collection costs and – as stated above – may even induce false reports of visits that actually did not occur. To minimize the respondent burden, data collection costs and measurement error, we need to minimize false positives.

A second malfunction would be when no survey is triggered, even though the geofence was visited (false negative). If the app fails to trigger a survey, even though the participant visited the geofence, we fail to cover part of the events of interest. This will lead to an underestimation of the frequency of such visits and decrease the statistical power for analyzing such visits. Furthermore, if false negatives are systematically related to the attributes of the visit (e.g., duration), they might potentially bias the estimates of any statistics produced from the geofence surveys.

Unfortunately, our design does not allow us to examine false negatives. Having a geolocation measurement at best every 30 minutes (see Bähr et al. (2020) for reasons why the intervals might have been longer), we are never able to verify whether a participant remained within the fence between two measurements. To be able to verify whether a participant remained within the fence between two measurements, we would need a higher frequency of geolocation measurements.

We can, however, approximately evaluate the false positives by comparing the geofence survey trigger via the Google Geofence API with our custom geolocation measurement.

The Google Geofence API was programmed to trigger a survey after identifying that the participant dwelled for 25 minutes within the geofence and then exited the geofence.

Dwelling within the geofence means that the geolocation along with its location uncertainty is within the geofence. It is very unlikely that this condition was fulfilled if none of our geolocation measurements designed to be taken in 30 minute intervals lies within the geofence.

To compare the Google Geofence API and our custom function, we use an explorative approach by creating a figure for each triggered survey. Figure 5.4 shows how it looks in our data when the geofence triggered survey matches our custom function. The x-axis in Figure 5.4 shows the time of the day that the geofence triggered a survey. The y-axis shows the distance of the participant to the job center. Each point represents a geomeasure from our customized function, which was designed to collect a geoposition every 30 minutes. The dashed line is the 200 meter mark for the geofence and the triangular shaped marker represents the triggered survey.

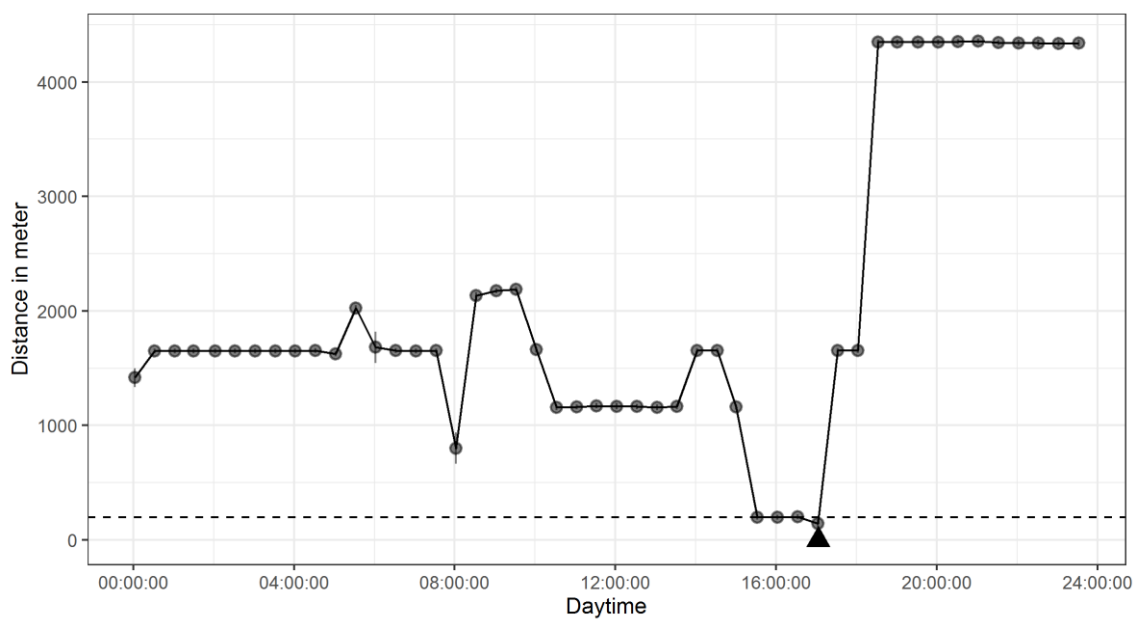


Figure 5.4: Example plot for the distances between the job center and the custom function geolocation measures on the day of the geofence triggered survey

During our explorative analysis, we find that in 121 out of 230 triggered surveys (52.6%) the Google API trigger matches our geolocation measure, similar to the example in Figure 5.4. In these cases, we are confident that the survey trigger using the geofence worked as intended. For the remaining 109 triggered surveys, we notice a pattern that deviates from that in Figure 5.4.

For 66 triggered surveys, we find that the location accuracy radius overlaps the area inside and outside of the geofence (i.e., participants could have been within the geofence or not). We assume that the geofence survey trigger also worked correctly for those cases but that the different measurement time points and possibly different accuracies between the Google Geofence API and our custom function lead to those mismatches.

We find 28 triggered surveys in which participants were not within the geofence prior to the survey trigger. We have no explanation for why this kind of mismatch appears.

For 15 triggered surveys, we do not have any geolocation measures from our custom function at least two hours prior to the time of the survey trigger. It may be possible that the Google API was able to collect geolocation data while our custom function was not. The lack of geomeasures from our custom function may be due to technical errors during the data transfer from the app to the backend or due to the Android operating system killing the data collection process (Bähr et al. 2020). Since the Google Geofence API was able to collect data, the Android operating system might discriminate between the custom data collection functions from third parties, like our IAB-SMART app, and functions developed and implemented by Google.

In addition to examining if participants visited the geofence prior to a triggered survey, some patterns indicate that a participant may not have visited the geofence for a job center visit but for another purpose. Figure 5.5, for example, shows a participant near the geofence from 7 am to 5 pm, which may indicate that the person works somewhere near

the job center, potentially even in the job center. It seems very unlikely but not impossible that this individual had a consulting meeting at the job center.

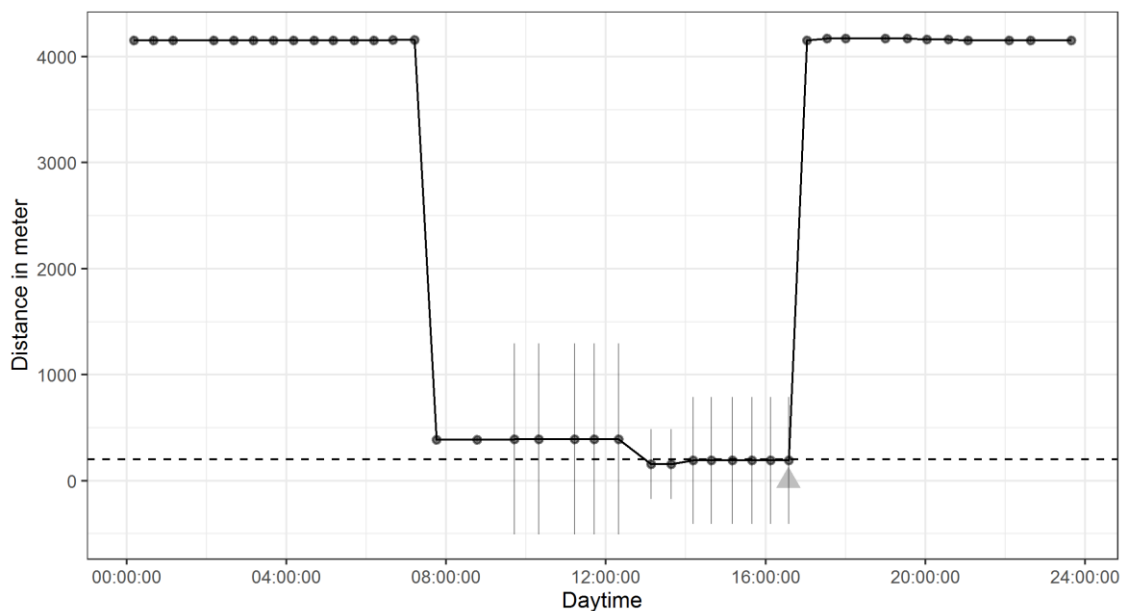


Figure 5.5: Example for distances between the job center and the custom function geo-location measures on the day of the geofence triggered survey

5.5 Conclusion - Lessons learned

In this paper, we described a geofence feasibility study with 410 geofences corresponding to job centers in Germany. Ideally, if an IAB-SMART participant visited a geofence with their smartphone for at least 25 minutes, a survey about the visit was triggered. In retrospect, we have to concede that many decisions we made were not optimal with respect to the data quality and data collection costs. Most errors we made originated from having no literature available on prior studies utilizing this data collection technique. We derive a series of lessons learned from our study that researchers may consider when designing and implementing a geofence survey in the data collection process. While some of these recommendations build on the specifics of our study design, population of interest and

research question about job center visits, they can inform researchers who plan to employ geofenced surveys in various contexts.

1. Collect information that indicates which geofence triggered a survey

When setting up the geofence surveys, we did not consider specifically instructing the programmers to save the information on which geofence triggered a survey. As a result, this exact information was lost, and we only know that one of the 410 geofences triggered a survey, but not which one. For an evaluation of how well the geofence surveys worked or to compare estimates between job centers, we needed to infer which geofence triggered a survey from a different, unrelated function in the app. Especially, studies that use more than one geofence should make sure to program the information on which geofence triggered a survey into their app.

2. Avoid overlapping geofences

If geofences overlap and an individual is within the overlapping space, a triggered survey cannot be reliably assigned to the geofence since one of four possible scenarios happens.

	Individual visits geofence A...	Individual visits geofence B...
...geofence A triggers a survey	correct trigger	wrong trigger
...geofence B triggers a survey	wrong trigger	correct trigger

Therefore, researchers should avoid overlapping geofences by, for example, creating more precise geofences around the locations of interest, such as our job center, that include just a small buffer of a few meters instead of using a simple large radius of several

hundred meters as we did. Such a more sophisticated approach would increase the time and effort needed when defining each geofence individually, but it would increase precision by avoiding overlaps and reducing the chance of a survey being triggered by a passerby, which might be especially problematic in densely frequented areas.

3. Consider the operation times of the locations in the geofence

As a result of not considering any operation times of the job centers, we find false triggered surveys in our data (i.e., triggered surveys on weekends and after business hours). Not considering operation times may not only increase the number of false triggered surveys but also the number of false reports. Even if the exact operation times are not known or they vary between geofences, it may be a good strategy to define broad operation times for all geofences to single out at least part of the false triggered surveys. Especially when the researchers are interested in locations that operate at certain hours only, geofences should be implemented with operation times to prevent false triggered surveys and reduce the amount of data cleaning required afterwards.

4. Consider the number of valid geofences per participant

In our design, each participant was able to trigger a survey for each geofence. In practice, however, each participant will only have one job center that is responsible for them. Therefore, we have participants who triggered a survey for a geofence of a job center that is not responsible for them. In cases where valid geofences differ between respondents, we therefore propose to link each participant to their valid geofence(s) (e.g., via the home address of the participant).

5. Availability of the survey invitation

As it is a major benefit of a geofence survey to collect real-time feedback, researchers should consider for how long a geofence survey should be available to participants. If an individual responds to a geofence survey after a day, survey questions may be out of context or the individual may have a harder time recalling events.

6. Validate the geofence visit and the event

We assumed that not every participant who visited the job center geofence was there for a consulting meeting. Therefore, we implemented a question in the geofence survey that asked if a consulting meeting took place or not. Participants who reported not being in the geofence for a consulting meeting were filtered out and did not receive the follow-up questions about the job center visit. From our design, we cannot validate whether a participant answering “no” to the validation question visited the job center at all or for a different purpose (i.e., we cannot distinguish between administrative visits to the job center and visits to other locations within the geofence). To distinguish between these cases, we should have implemented two validation questions: one that asks whether the participant was at the point of interest, and a second one that asks about the specific purpose of the visit (e.g., a consulting meeting). Researchers need to consider the context in which their study is conducted to determine what questions need to be asked to validate whether a survey is triggered in the appropriate context (i.e., at the right time, at the right location, and for the right person).

5.6 Use of geofences in future research

As noted earlier, the lessons learned from the implementation of geofences are informed by the scope of our IAB-SMART study, and not all recommendations might apply to all

geofence surveys in other contexts. To broaden our understanding of when geofencing can be used as a valuable tool in survey research, we need more studies that implement the technology in the data collection process and validate the findings in different settings. Based on our experience, working with 410 geofenced job centers might have been too ambitious of a task. There still seem to be many technical and logistic problems pertaining to the accuracy of the geoposition measurement and the validation of locations to simultaneously implement several hundred geofences in one study. For example, while it is highly unlikely that a job center visit happened at 10 on a Saturday evening, our definition of opening hours between 7am and 7pm for all job centers might have been too imprecise. However, for studies with one or just a few precisely defined geofences, such as a stadium where the spectators of a sports event or concert visitors should be invited to an experience survey, this technology could be a very promising addition to the toolkit of survey designers.

References

- Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., and Trappmann, M. (2020). Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data. *Social Science Computer Review*. <https://doi.org/10.1177/0894439320944118>.
- Bradburn, N. (1978). Respondent burden. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 35-40.
- Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R. and Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78(3), 721-733.
- Grennwood, A. (2017). The Importance of Geofencing Market Research Surveys. [Blog Post]. Available at: <https://blog.flexmr.net/the-importance-of-geofencing-market-research>. (Accessed January 2020).
- Haas, G.-C., Kreuter, F., Keusch, F., Trappmann, M. and Bähr, S. (2020). Effects of Incentives in Smartphone Data Collection. In *Big Data Meets Survey Science* (eds C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japac, A. Kirchner, S. Kolenikov and L.E. Lyberg). doi:10.1002/9781118976357.ch13
- Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., and Trappmann, M. (2020a). Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a German national survey. *Sociological Methods & Research*. Published online before print April 7, 2020. DOI: 10.1177/0049124120914924
- Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., and Trappmann, M. (2020b). Participation rates and bias in a smartphone study collecting self-reports and passive mobile

- measurements using a research app. *Paper presented at AAPOR 75th Annual Conference*, Virtual Conference, June 11-12.
- Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., and Trappmann, M. (2018). Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent. *Social Science Computer Review*. Published online before print December 18, 2018. DOI: 10.1177/0894439318816389.
- Stone, A. A., and Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, 16(3), 199–202.
<https://doi.org/10.1093/abm/16.3.199>.
- Trappmann, M., Bähr, S., Beste, J., Eberl, A., Frodermann, C., Gundert, S., Schwarz, S., Teichler, N., Unger, S. and Wenzig, C. (2019). Data Resource Profile: Panel Study Labour Market and Social Security (PASS). *International Journal of Epidemiology*, 2019, 1411–1411g.
- Trappmann, M., Müller, G., and Bethmann, A. (2013). Design of the study. In: User guide “Panel Study Labour Market and Social Security” (PASS): Wave 6, edited by A. Bethmann, B. Fuchs, and A. Wurdack, 13–22. Available at:
http://doku.iab.de/fdz/reporte/2013/DR_07-13.pdf.
- Tourangeau, R., Rips, L.J., and Rasinski, K.A. (2000): *The Psychology of Survey Response*. Cambridge University Press.
- Wray, T. B., Pérez, A. E., Celio, M. A., Carr, D. J., Adia, A. C., and 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(5), 900–906.

<https://doi.org/10.1111/acer.13991>.

Appendix

Question Number

Screenshot

English Translation

2

Im Folgenden geht es um Ihre persönliche Erfahrung, die Sie HEUTE mit dem Jobcenter und den Mitarbeitern gemacht haben. Inwieweit treffen aus Ihrer ganz persönlichen Sicht die folgenden Aussagen zu?

Die Mitarbeiter des Jobcenters haben mich HEUTE bevormundet.

☐ Trifft voll und ganz zu

☐ Trifft eher zu

☐ Trifft eher nicht zu

☐ Trifft überhaupt nicht zu

BACK CONTINUE

The following is about your personal experience with the job center and its employees that you have made TODAY. To what extent do the following statements apply from your very personal point of view?

The employees of the job center patronized me TODAY.

- ☐ Strongly agree
- ☐ Slightly agree
- ☐ Slightly disagree
- ☐ Strongly disagree

3

Im Folgenden geht es um Ihre persönliche Erfahrung, die Sie HEUTE mit dem Jobcenter und den Mitarbeitern gemacht haben. Inwieweit treffen aus Ihrer ganz persönlichen Sicht die folgenden Aussagen zu?

Ich hatte HEUTE keine Möglichkeit, meine eigenen Vorstellungen in dem Gespräch einzubringen.

☐ Trifft voll und ganz zu

☐ Trifft eher zu

☐ Trifft eher nicht zu

☐ Trifft überhaupt nicht zu

BACK CONTINUE

The following is about your personal experience with the job center and its employees that you have made TODAY. To what extent do the following statements apply from your very personal point of view?

TODAY I had no opportunity to bring my own ideas into the conversation.

- ☐ Strongly agree
- ☐ Slightly agree
- ☐ Slightly disagree
- ☐ Strongly disagree

4

Im Folgenden geht es um Ihre persönliche Erfahrung, die Sie HEUTE mit dem Jobcenter und den Mitarbeitern gemacht haben. Inwieweit treffen aus Ihrer ganz persönlichen Sicht die folgenden Aussagen zu?

Die Mitarbeiter des Jobcenters haben mit mir HEUTE ausführlich besprochen, wie ich meine Chancen auf dem Arbeitsmarkt verbessern kann.

☐ Trifft voll und ganz zu
☐ Trifft eher zu
☐ Trifft eher nicht zu
☐ Trifft überhaupt nicht zu

BACK CONTINUE

The following is about your personal experience with the job center and its employees that you have made TODAY. To what extent do the following statements apply from your very personal point of view?

The employees of the job center have discussed with me TODAY in detail how I can improve my chances on the job market.

- ☐ Strongly agree
- ☐ Slightly agree
- ☐ Slightly disagree
- ☐ Strongly disagree

5

Im Folgenden geht es um Ihre persönliche Erfahrung, die Sie HEUTE mit dem Jobcenter und den Mitarbeitern gemacht haben. Inwieweit treffen aus Ihrer ganz persönlichen Sicht die folgenden Aussagen zu?

Ich hatte HEUTE das Gefühl, dass ich den Mitarbeitern vertrauen kann.

☐ Trifft voll und ganz zu
☐ Trifft eher zu
☐ Trifft eher nicht zu
☐ Trifft überhaupt nicht zu

BACK CONTINUE

The following is about your personal experience with the job center and its employees that you have made TODAY. To what extent do the following statements apply from your very personal point of view?

TODAY I had the feeling that I can trust the employees.

- ☐ Strongly agree
- ☐ Slightly agree
- ☐ Slightly disagree
- ☐ Strongly disagree

6

Im Folgenden geht es um Ihre persönliche Erfahrung, die Sie HEUTE mit dem Jobcenter und den Mitarbeitern gemacht haben. Inwieweit treffen aus Ihrer ganz persönlichen Sicht die folgenden Aussagen zu?

Es wurden HEUTE nur Forderungen gestellt, statt mir wirklich zu helfen.

☐ Trifft voll und ganz zu

☐ Trifft eher zu

☐ Trifft eher nicht zu

☐ Trifft überhaupt nicht zu

BACK CONTINUE

The following is about your personal experience with the job center and its employees that you have made TODAY. To what extent do the following statements apply from your very personal point of view?

There were only demands made TODAY instead of really helping me.

- ☐ Strongly agree
- ☐ Slightly agree
- ☐ Slightly disagree
- ☐ Strongly disagree

7

Im Folgenden geht es um Ihre persönliche Erfahrung, die Sie HEUTE mit dem Jobcenter und den Mitarbeitern gemacht haben. Inwieweit treffen aus Ihrer ganz persönlichen Sicht die folgenden Aussagen zu?

Die Mitarbeiter des Jobcenters haben mir HEUTE geholfen, eine neue Perspektive zu entwickeln.

☐ Trifft voll und ganz zu

☐ Trifft eher zu

☐ Trifft eher nicht zu

☐ Trifft überhaupt nicht zu

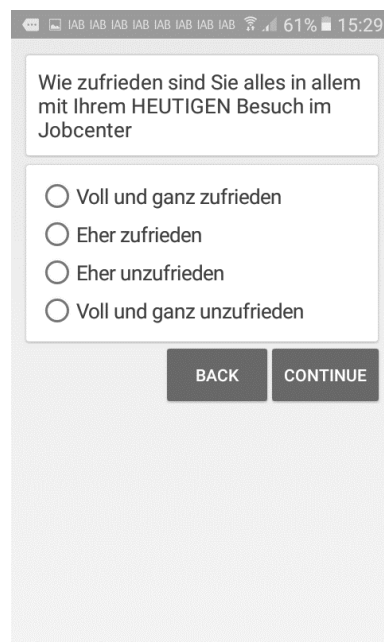
BACK CONTINUE

The following is about your personal experience with the job center and its employees that you have made TODAY. To what extent do the following statements apply from your very personal point of view?

TODAY the employees of the Job Center helped me to develop a new perspective.

- ☐ Strongly agree
- ☐ Slightly agree
- ☐ Slightly disagree
- ☐ Strongly disagree

8



Wie zufrieden sind Sie alles in allem mit Ihrem HEUTIGEN Besuch im Jobcenter

☐ Voll und ganz zufrieden

☐ Eher zufrieden

☐ Eher unzufrieden

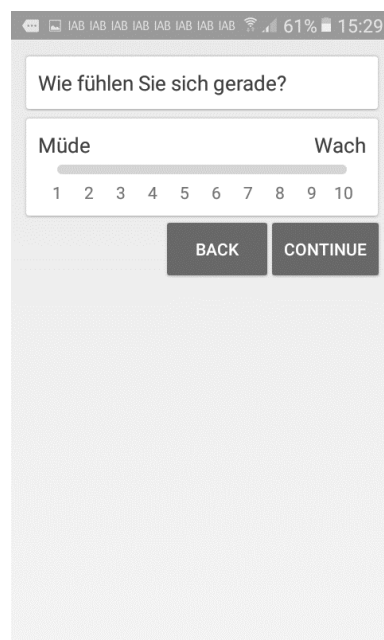
☐ Voll und ganz unzufrieden

BACK CONTINUE

How satisfied are you all and everything with your TODAY visit to the job center

- ☐ Fully satisfied
- ☐ Slightly satisfied
- ☐ Slightly dissatisfied
- ☐ Fully dissatisfied

9



Wie fühlen Sie sich gerade?

Müde Wach

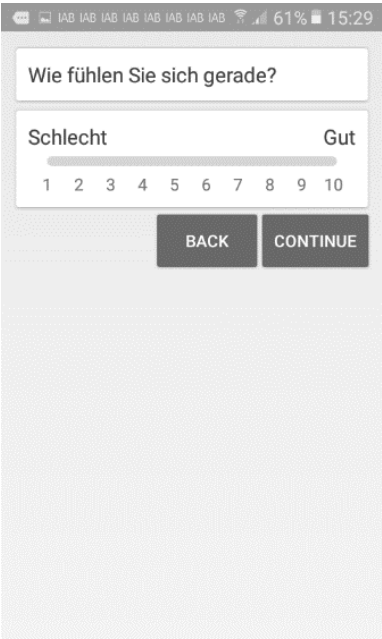
1 2 3 4 5 6 7 8 9 10

BACK CONTINUE

How are you feeling right now?

Tired – Awake

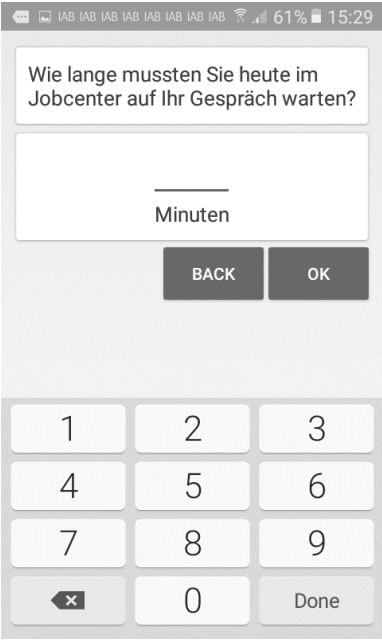
10



How are you feeling right now?

Bad – Good

11



How long did you have to wait for your interview at the job center today?

Minutes

Figure 5.6: Screenshots and English translation of the geofence survey for IAB-SMART participants that received the survey invitation and answered the first question with “Yes”.

6 Conclusion

Over the last decade, information and communication technology lead to tremendous changes in society, resulting in challenges and opportunities for the survey profession. The web mode, smartphones and the combination of a seemingly unmanageable variety of different data sources offer new possibilities to design and modernize data collection approaches that need to be evaluated in their feasibility and effects on data quality. In my thesis, I tackled some of these new possibilities to modernize the design of data collection methods by evaluating different research designs in each of my four thesis papers. My thesis papers originate from three different projects, consider different target populations (i.e., German establishment population, Egyptian parents, general German population), and different modes (web survey, text message survey, app based data collection). Three of the four submitted papers use an experimental design to reach their conclusion. The fourth paper reports on a novel approach to use organic or sensor data to trigger surveys at certain locations. On the first sight, the papers of my thesis may miss to hit the same notch. However, all four papers have in common that they use self-administered data collection tools, i.e., for the data collection, no interviewers were used but individuals provided their data themselves and use novel approaches in their respective research field. By showing how the data collection process was designed, all four submitted papers make necessary contributions to the academic research literature on modernizing data collection methods and enlarge the survey methodologist's toolbox for similar studies.

My first thesis paper, *Comparing Response Burden between Paper and Web Modes in Establishment Surveys*, evaluates the difference in response burden between a paper and web mode in a German establishment survey. Response burden was measured with three variables (estimated time to complete the questionnaire, perceived time and burden). To

evaluate if response burden is lower in an establishment web survey and whether respondents feel less burdened if they can choose between a paper and a web mode, four mode comparisons were made (*Paper-only* vs. *Web-only*, *Choice-Paper* vs. *Paper-only*, *Choice-Web* vs. *Web-only*, *Choice-Paper* vs. *Choice-Web*). Furthermore, within each mode group establishments were randomly assigned to two different topics. The response burden study shows that web respondents, whether they were offered web as a standalone mode or concurrently with a paper questionnaire, have no negative effect on response burden and a small positive effect on the estimated time to fill out the questionnaire. As there are no substantial differences between the paper and web mode design, the results suggest that the web mode is a suitable alternative or add-on for establishment surveys that already use a paper mode.

My second thesis paper, *Comparing Single-sitting Versus Modular Text Message Surveys in Egypt*, investigates the effects of two designs to administer text message surveys: single-sitting and modular and contribute to the knowledge on how to implement text message surveys. Overall, 1,081 Egyptian parents were randomly assigned to one of both groups. While the single-sitting group received one invitation to an eight-question long text message survey, the *modular* group received an invitation to a question each day over the course of eight days. Results show that compared to the single-sitting design, the modular design achieved a higher number of answered questions but had fewer fully completed questionnaires. Furthermore, the paper finds some differences in substantive responses of behavioral questions between the groups. The study suggests no differences in nonresponse bias between both groups and in the probability to respond to a follow-up survey.

My third thesis paper, *Effects of Incentives in Smartphone Data Collection*, investigates the effects of monetary incentives on participation in a smartphone study. Incentives are

one of the main tools to increase participation and cooperation in a survey. However, smartphone studies that collect sensor data may be seen as more valuable by participants and incentives may be higher than in surveys. Furthermore, smartphone studies enable new strategies to keep participants engaged, that is, keeping the app installed. To make a first step towards an effective incentive strategy for smartphone studies, the study investigates a crossed two factor experimental design. First, participants were promised either 10 or 20 Euros conditional on installing the app. Second, participants were promised either one Euro for each passive data collection function that was activated for 30 consecutive days or one Euro per function plus a five Euro bonus if all five data sharing functions were activated for 30 consecutive days.

The amount of incentive offered for installing the app (10 Euro vs 20 Euro) influences the installation rate. Compared to a 10-Euro incentive, individuals that were offered 20-Euro install the app more often (13.1% vs. 16.4%). However, the study shows no evidence that installation incentives affect the number of activated functions, number of deactivated functions or retention. The second experiment, paying respondents a five-Euro bonus incentive if they grant access to all five data sharing functions does not affect the propensity to install the app. As the bonus incentives doubles the monthly incentive when all functions are activated, one would expect a substantial effect on the propensity to activate functions, keep functions activated and retention. In this regard, however, the study shows no evidence that there are differences between experimental groups. Combining the two experiments from the crossed two factor experimental design and adding the planned amount of 20 Euro for responding to in app surveys, we get four different maximum amounts of incentives for that study participants could earn: 60, 70, 90 and 100 Euro. The study shows no evidence that this promised maximum amount affects the de-

cision to install the app, activate more or less functions or deactivate any functions. However, participants in the 70 Euro and above group kept the app installed for a longer period.

One of the major ethical concerns conducting this study was that the offered incentive is too high for vulnerable groups, as they may feel forced to install the app and keep it installed. The study found no difference in installation rates, number of activated functions, number of deactivated functions or retention indicating, that experimental incentive groups have an effect on the vulnerable groups in our study. However, not seeing any differences between welfare recipients and non-welfare recipients does not mean that particular individuals did not feel forced to participate in the study as the situation of a particular welfare recipient does not allow to decline the offered incentive.

My fourth thesis paper, *Using Geofences to Collect Survey Data: Lessons Learned From the IAB-SMART Study*, evaluated the feasibility of geofences for labor market research. In context of survey research a geofence can be defined as a geographical area that triggers a survey by entering this area, dwelling within this area for a defined amount of time and/or exiting this area. The design approach in this study combines survey and smartphone sensor data by using geolocation data to trigger survey invitations. For this purpose, we implemented the geolocation of 410 German job center with a 200 m radius defining the geofences in the IAB-SMART app. If participants dwelled for 25 minutes within one of the geofences, a survey invitation was sent, containing questions about the experience of a consultation meeting in the job center.

By now, there is a vast body of literature on how to design and implement surveys in different settings. However, there are only few articles that use geofences and a vast body of literature about the “do’s and don’ts” does not exist. Without any background on how

to design a geofence survey, design decisions were not optimal for this study. The submitted paper describes design decisions made and concludes with six lessons learned on how to improve the design: (1) Collect information that indicates which geofence triggered a survey, (2) Avoid overlapping geofences, (3) Consider the operation times of the locations in the geofence, (4) Consider the number of valid geofences per participant, (5) Availability of the survey invitation, (6) Validate the geofence visit and the event. Not all recommendations may apply to other geofence studies and the study did not cover all design decisions. However, this is a first step towards designing survey data collection with geolocation data.

The insights gained from my thesis may be of assistance to researchers designing data collection tools in different contexts. Each paper in this thesis contributes to a growing body of literature and has different outcome variables. The response burden study, for example, provides important findings for the development and design of web establishment surveys. Results from the text message survey study contribute to a growing body of literature on how to apply this mode. The incentive study, from the IAB-SMART app, provides first insights on how different incentive strategies affect the participation on smartphone studies. And the geofence study, which uses a novel approach to administer surveys, provides a guideline to avoid several design flaws when setting up a geofence study.

Eidesstattliche Versicherung

Eidesstattliche Versicherung gemäß § 9 Absatz 1 Buchstabe e) der Promotionsordnung der Universität Mannheim zur Erlangung des Doktorgrades der Sozialwissenschaften:

1. Bei der eingereichten Dissertation mit dem Titel „Modernization of Data Collection Methods“. handelt es sich um mein eigenständig erstelltes eigenes Werk.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bisher nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikations-leistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erkläre und nichts verschwiegen habe.

Georg-Christoph Haas