sciendo

Journal of Official Statistics, Vol. 38, No. 4, 2022, pp. 1069–1095, http://dx.doi.org/10.2478/JOS-2022-0046

The Effects of Response Burden – Collecting Life History Data in a Self-Administered Mixed-Device Survey

Johann Carstensen[1], *Sebastian Lang*[1], *and Fine Cordua*[1]

Collecting life history data is highly demanding and therefore prone to error since respondents must retrieve and provide extensive complex information. Research has shown that response burden is an important factor influencing data quality. We examine whether increases in different measures of response burden in a (mixed-device) online survey lead to adverse effects on the data quality and whether these effects vary by the type of device used (mobile versus non-mobile).

We conducted an experimental study in an online mixed-device survey, for which we developed a questionnaire on the educational and occupational trajectories of secondary-school graduates, undergraduates, and university graduates. To address our research question, we randomly assigned different levels of response burden to the participants and compared different measures on the data quality and response.

We found mixed evidence for unfavourable effects of response burden on the examined outcomes. While some of our results were expected, they were not consistent across all subgroups. Most interestingly, the effects of response burden on outcomes seemed to differ based on the device used. Hence, we conclude that further research is needed to optimise the collection of complex data from different groups of participants.

Key words: Respondent burden; data quality; response behavior; experiment; online survey.

1. Introduction

In attempting to reduce bias in survey results, researchers have increasingly turned their attention to response or respondent burden. Response burden can introduce bias through unit (Groves et al. 2011) or item nonresponse (Galesic and Bosnjak 2009), as well as through measurement error (Peytchev and Peytcheva 2017). Thus, survey design to reduce response burden must be carefully balanced with the substantial needs of a data collection endeavour.

Recently, many survey-makers have strived to collect increasingly complex information using web surveys as the most cost-effective survey mode. However, collecting complex life history data – that is, detailed biographical data suited for complex analysis using for instance event history models – in a self administered mode poses new challenges to minimising measurement error as there is no interviewer who can provide help and motivation. One of these challenges concerns reducing response burden.

[1] German Centre for Higher Education Research and Science Studies, Lange Laube 12, 30159 Hannover, Germany. Emails: carstensen@dzhw.eu, lang@dzhw.eu and cordua@dzhw.eu

Acknowledgments: This research was supported by the German Federal Ministry of Education and Research (16P4246). The authors would like to thank the editors and anonymous reviewers as well as the participants of the 8th Conference of the European Survey Research Association and the joint meeting of the Association of Social Science Institutes (ASI) and the DGS section "Methods of Empirical Social Research" for helpful comments on earlier versions of this article.

© Statistics Sweden

Multiple factors can influence response burden, such as varying the length or frequency of a survey or the amount and complexity of the requested information. In this experimental study, we vary both the amount and complexity of items regarding life history, to determine how different measures of data quality are influenced by those variations. Using a web survey, we also distinguish survey respondents who use mobile devices from those using desktop computers, to identify heterogeneous effects.

We start by elaborating on our definition of response burden. This serves as the basis to introduce Tourangeau's (2000, 2018) response process approach, which can explain why participants experience various difficulties when answering survey questions. Additionally, satisficing theory (Krosnick 1991) is used to explain participants' reactions on an increasing response burden when engaged in reporting. We then proceed into a review of recent research on the factors that foster response burden. The methods used and the study design are then combined with information from recent research to develop testable hypotheses. We reveal our strategy and describe our data (Section 2) before presenting our results (Section 3). We conclude with a discussion and summary (Section 4).

1.1. Theory

According to Bradburn (1979, 36), response burden' is not an objective characteristic of the task, but is the product of an interaction between the nature of the task and the way in which it is perceived by the respondent.' He identified four key elements of respondent burden (Bradburn 1979, 36): (1) the length of the interview; (2) the amount of effort required of the respondent; (3) the amount of stress on the respondent; and (4) the frequency with which the respondent is interviewed. Three of these elements can be seen as properties of a survey, while the amount of stress on the respondent is a product of the interaction between the survey contents and characteristics of the respondent. While more recent research has often relied on a respondent-centred approach in defining response burden (Hedlin et al. 2008), we focus on properties of a survey, varying them experimentally.

To understand the impact of response burden on the response process, response rates, and data quality, we refer to Tourangeau's (2000, 2018) cognitive response process model. Based on Tourangeau (2000, 2018), participants' response process when answering a survey can be summarised in four steps. First, respondents must **comprehend** the posed question in a way that fits the researchers' conceived concept. Second, they must **retrieve** information from their memory that enables them to answer the question. Third, **judgement and estimation** take place: the respondents can estimate an answer (e.g. the frequency of an event in a specific time period), count, rate, or rely on existing information. Fourth, the respondents **report** the answer. Regarding response burden – as defined above – error may be introduced in this process in different manners.

While **comprehension** error is, by definition, connected to the (subjective) difficulty of a task, it is also more likely to occur due to fatigue in longer interviews and when more detailed information is asked. This is true for all four steps of the cognitive process. Moreover, when increasing the amount of detail in a survey, it becomes necessary to be increasingly specific in formulating the question or instruction. This consequently increases the likelihood that the researchers' intended meaning and the respondents' understanding of a question do not match. Specifically in relation to collecting life history

data, this can happen, for example, in identifying timespan or granularity ("daily", "monthly", "yearly") in reporting events.

Retrieval may be hampered when respondents must recall less salient information, for example when having to report detailed information on long time spans. Fatigue may furthermore increase retrieval errors. Such errors may occur due to forgetting or forward telescoping.

The same general arguments apply to the third step, **judgement and estimation**: the probability of errors is also likely to increase here if more, very detailed, or difficult information is requested, or if this information lies further in the past, in which case respondents are more likely to have to guess or estimate their answers instead of clearly recalling details. Respondents' strategies in this step may vary between low and high effort, depending on the time they have (Burton and Blair 1991). Besides the tendency that available time will decrease in a longer survey, fatigue may lead to choosing low-effort strategies.

Finally, respondents may introduce error in the fourth step, **reporting**. Again, in addition to effects through fatigue, requesting more (and more detailed) information or asking more difficult questions should increase the probability of errors in reporting, as the task of mapping information to scales provided by researchers becomes more complex. In collecting life history data, this could occur when researchers ask, for instance, about changes in the kind of employment contract and require respondents to be more specific about the types of such contracts than they normally would be; this could represent a mismatch between the scale provided and the information respondents would be able to remember.

1.2. State of Research

The *length of the survey* (Bradburn 1979) and the *amount of detail* asked (Warriner 1991) can complicate the process of retrieval because the respondents need to concentrate for a longer time (Crawford et al. 2001; Galesic 2006; Galesic and Bosnjak 2009; Marcus et al. 2007). Relatedly, the *time needed* to participate is found as a factor that increases response burden (Crawford et al. 2001; Galesic 2006; Galesic and Bosnjak 2009; Marcus et al. 2007) and self-reported response burden increases with interview duration (Guin et al. 2012; Yan et al. 2020). The *reference period* has a major influence on participants' response burden. The farther back in the past an event lies, the more poorly it is likely to be remembered (Groves et al. 2011; Tourangeau 2000), and the more uncertainty it can create, for instance when answering questions requires comprehensive processes of recalling memories (Drasch and Matthes 2013). Relatedly, the *difficulty of the item* is another element of the response burden (Galesic and Bosnjak 2009; Hambleton et al. 1991; Reise et al., 2005; Zanon et al. 2016). The *more difficult* it is for the participants to read and understand an item, the *more effort is required to retrieve the information demanded and judge or estimate an answer*, and the lower the participants' *cognitive capacity* is, the higher their response burden is likely to be (Guin et al. 2012; Gummer and Roßmann 2015; Lenzner 2012; Tourangeau 2003; Yan and Tourangeau 2008).

Questions of higher difficulty can lead to difficulties in retrieval (e.g. forgetting events), and can thus produce errors in reporting (Tourangeau 2000). Such errors can also appear due to respondents' tactics to minimise their response burden. Recent research fruitfully explains respondents' reactions to high response burden by applying decision field theory

(e.g. Galesic 2006). 'When confronted with a difficult personal decision, the decision maker tries to anticipate and evaluate all of the possible consequences produced by each course of action.' (Busemeyer and Townsend 1993, 444). Answering questions in a survey can be seen as such a 'difficult personal decision'.

To avoid high effort, respondents may react to a high response burden by *satisficing*. As satisficing theory states, this may lead to incomplete or biased information. Strategies may include wrongly agreeing or disagreeing with statements (e.g. by choosing the first answer regardless of its content), avoiding socially critical statements, or giving no answer (e.g. by choosing the category 'I don't know') (Krosnick 1991). As another means to reduce effort, participants may react with *underreporting frequencies* of events (Kaper and Saris 1999) and thus – more or less intentionally – answering questions in a way that shortens the interview (Peytchev and Peytcheva 2017).

These results seem to be especially relevant when deciding how to go about collecting biographical information. For instance, *looping questions* are questions that are asked several times in the course of the survey and refer to recurring events (e.g. gainful employment), where the frequency of repetition depends on the frequency of the event. This sort of questions is often used in collecting life history data and is also implemented in the questionnaire on which this study is based. Depending on their implementation, looping questions can either lead to inaccurate numbers in event reports but detailed information on the reported events (when asking for subsequent events after each report of the preceding event) or to accurate numbers in event reports but missing information on the details of the reported events (when first asking for the frequency k of an event and then asking for details on each of the reported k events) (Eckman and Kreuter 2018).

Furthermore, respondents can react by not answering parts of the questionnaire (*item-nonresponse*, see Galesic and Bosnjak 2009) or by refusing to answer the interview completely (*unit-nonresponse,* see Groves et al. 2011). Finally, a burdensome survey is discussed to cause *panel attrition*, meaning that it reduces the participant's willingness to participate in future surveys, for instance because the preceding interview experience was frustrating (Hoogendoorn and Sikke 1998; Sharp and Frankel 1983).

Findings on the *device* participants used to answer web surveys have been inconsistent. Some studies have found no evidence for the assumption that mobile, compared to non-mobile devices, could lead to a higher response burden concerning dropout, item nonresponse, the time needed to complete the survey, or the amount of detail an answer contained (Antoun et al. 2017; De Bruijne and Wijnant 2014; Marcus et al. 2007; Toepoel and Lugtig 2014). Others have found no tendencies for respondents to agree regardless of the content (acquiescence), to provide non-substantive answers (congruent with Mavletova 2013), to choose midpoint values or the first response option (primacy effects; congruent with Mavletova 2013), or to choose answers depicted at the same position in the survey (straightlining; Clement et al. 2020), when using mobile devices. Moreover, Daikeler et al. (2020) have not identified more misreporting and Krebs and Höhne (2020) have not identified stronger scale direction effects, when mobile devices have been used. At the same time, researchers have observed that respondents who used mobile phones in a previous survey tended not to use them in the next survey they took (Lugtig and Toepoel 2016). Additionally, they produced more missing data (Keusch and Yan 2017) and showed lower response rates (Lee et al. 2019; Mavletova 2013). This could indicate a tendency for

respondents using mobile devices to experience a higher response burden (Buskirk and Andrus 2014).

Apart from one study that reported lower response time for mobile users (Lee et al. 2019), studies were found to report higher durations in survey completion when respondents used smartphones, compared to stationary computers (Antoun et al. 2017; De Bruijne and Wijnant 2014; Schlosser and Mays 2018). This could be partly ascribed to the need for additional scrolling (Couper and Peterson 2017). The time needed to answer on mobile devices seemed to be less pronounced for respondents who were familiar with them (Antoun and Cernat 2020). Also, participants using mobile devices were more likely to be outside their homes and among other people, restricting the attention they paid to the survey (Antoun et al. 2017; De Bruijne and Oudejans 2015).

Collecting life history data is associated with especially high response burden, as all the above mentioned factors are involved. Posing questions towards several episodes in the respondents' lives seems to inevitably lead to longer surveys than those limited to particular episodes, as the reference period tends to cover events that lie farther back in the respondents' memories. Also, the amount of detail asked tends to be high, as connections between the reported episodes are enquired. Answering questions that encompass life history data can thus turn out to be rated as difficult for the respondents. Such factors likely lead to an increase in the time needed to respond to life history questions.

Starting from education and leading up to the labour market, career paths are highly dependent on preceding episodes in life. Hence, collecting data within a life history framework seems especially promising to gain a comprehensive understanding of how students decide for or against different paths. With this in mind, the present article investigates response burden when collecting life history data from students and employees.

1.3. Hypotheses

In the following, we describe our study in the light of reseach gaps revealed by the literature. Thereafter, we derive our hypotheses.

First, we check whether our treatments have the expected effect on one indirect and one direct measure of response burden: response time and the subjective assessment of the questionnaire's difficulty (see H1a and H1b below). Especially the latter measurement of response burden has not been common in recent research. Many studies have relied on single variables, such as response time, as a measure for response burden only (e.g. Gummer and Roßmann 2015; Yan and Tourangeau 2008).

Second, we use several variables that measure the impact a high response burden should have on participants' behaviour (see H2a to H3c below). Conversely, many studies have focused on one effect response burden has on answering behaviour (e.g. Galesic 2006).

Third, only one study has focused on life history data. This study did not investigate response burden as such, but rather participants' recollections when filling out the survey (Drasch and Matthes 2013). Our study explicitly investigates the impact of response burden on life history data, since such data provides encompassing insights into respondents' processes of decision-making and behaviour. Additionally, we vary relevant variables that should impact response burden (such as the reference period and amount of detail asked; see Subsection 2.1 for more information).

Fourth, research has lacked considerations on participants' behaviour in self-administered web surveys, as many surveys were conducted when the internet was not as established in mainstream use. We offered participants in our survey a choice of which device to use (see H4 below). Altogether, this study provides important insights not only regarding the effects of response burden on life history data, but also regarding the possible moderation effects of devices on response burden.

In this article, we investigate the impact of response burden on several dimensions relating to response behaviour and data quality. We begin by analysing the effect of our experimental treatments on one indirect and one direct measure of response burden. These are response time (H1a) and the subjective evaluation of the difficulty of the questionnaire (H1b). Both should be higher when increasing response burden.

With respect to response behaviour, we use break-offs and panel consent as indicators. Based on our theoretical approach and the state of research presented in the previous section, increasing response burden should move respondents closer to or beyond their specific critical threshold for satisficing. Analogously, this should apply to the refusal of panel consent. Therefore, we assume that an increase in response burden leads to an increase in the probability of the interview being ended before completion (H2a) and a decrease in the probability that the respondent will consent to further interviews (H2b).

The same theoretical argument holds for data quality, with incorrect or insufficient information as other forms of satisficing. We must make some presumptions with respect to this, as we cannot exploit external data to verify the information the respondents provide, such as linked or linkable administrative data. These presumptions are as follows:

1. A higher share of months with no information on activities indicates poorer data quality,
2. A higher number of episodes indicates higher data quality, and
3. Less item non-response (INR) indicates higher data quality.

The simple argument behind all three presumptions is that more information is better. This applies especially for presumption (1), as the alternative would be no information on some periods. Regarding presumptions (2) and (3), we further argue that there is no plausible reaction to response burden, such as satisficing, that would lead to a higher number of episodes or less INR.

In general, we expect a decrease in the data quality with an increase in response burden. Based on these presumptions, we argue that a higher response burden should result in a lower number of episodes (H3a), a higher share of months with no information (H3b), and a higher number of INR (H3c).

Finally, we consider the heterogeneous state of research on mobile devices. Due to the inconsistent results of previous research on this topic, we expect heterogeneous effects depending on the respondents' choice of device. Overall, we think that arguments in favour of generally stronger effects of a higher response burden on mobile devices prevail. Therefore, we expect to find stronger effects of response burden for the participants who decided to answer the survey on a mobile device (H4).

2. Methods

2.1. Study Design and Data

This work is part of a project to prepare the Student Life Cycle Panel (SLC) of the German Centre for Higher Education Research and Science Studies (DZHW). For this project, we developed a questionnaire to collect life history data from secondary-school graduates, undergraduates, and university graduates in a self-administered web survey. As providing life history data is highly demanding for participants, the topic of response burden is of special interest in this context. To investigate the impact of response burden on participation, break-offs, and data quality, we conducted an experiment in which we varied response burden in the survey module for collecting life history data.

Our sample consists of four subsamples with separate sampling frames and sampling procedures: (1) (former) students, (2) school-leavers from vocational schools, (3) international students, and (4) non-traditional students.

Subsample (1) stems from a probability-based online access panel of (former) students maintained by the DZHW. Subsample (2) was drawn from vocational schools in North Rhine-Westphalia to include school-leavers qualified to enter higher education but who took vocational training or started to work immediately after leaving school. These respondents constitute an important comparison group for higher education research. Therefore, we paid special attention to ensuring the adequacy of our questionnaire for this group.

Subsamples (3) and (4) were recruited at a German university and represent two small subgroups with rather low participation rates in regular student samples. Like the members of subsample (2), they are of special interest for higher education research and could have specific requirements for the questionnaire.

For our analyses, we pooled these subsamples and controlled for the origin of the observations.

The participants were invited by e-mail, and received two reminders (each at intervals of three to four weeks). Participation was incentivised with a lottery of one big prize (Apple iPad Pro, 10.5″, 64GB), plus 2 X EUR 100, and 10 X EUR 50-book vouchers. The field phase ended on January 17, 2019. The overall response rates were 8.4% (RR1) and 8.6% (RR2; American Association for Public Opinion Research 2016).

The questionnaire consisted of three main parts. The first and the last part contained about 50 questions on socio-demographics and the respondents' current activities. The middle part was the life history module, where the experimental treatments were implemented. Two experimental variations of the response burden were made in a fully-crossed, 2 X 3 design. (For an overview on the distribution of observations in experimental treatments and types of devices, see Figure 1.)

First, the reference period for which life events should be reported was varied between two and five years. Second, the amount of detail the respondents were asked to provide in the module on dependent employment was varied. We implemented a minimum, medium, and maximum information condition, where the request for additional information in these modules should represent an incremental increase in the response burden, as the medium and maximum condition require more information, each with an increasingly high resolution. (For further details about the information asked in each version, see Table 1.)

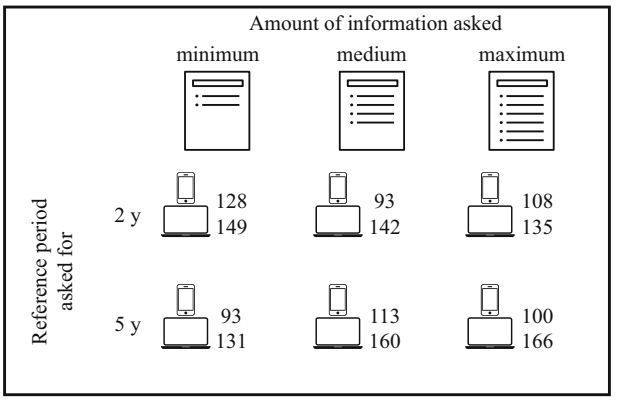


Fig. 1. Number of respondents by experimental treatment and device used to complete the survey.
Note: The experimental treatments (i.e. reference period asked for and the amount of information asked for) were randomly assigned. Device used to complete the survey (i.e., mobile versus non-mobile) was not randomly assigned.

Table 1. Information asked in minimum, medium, and maximum information condition.

	Minimum information condition	Medium information condition	Maximum information condition
Criteria splitting episodes into sub-episodes	- employer - occupation - occupational status - place of work	- employer - occupation - occupational status - place of work	- employer - occupation - occupational status - place of work
Additional information (asked chronologically for existing episodes)	- temporary contract (yes/no) - employment contract - working time (only starting month) - income (only starting month) - economic sector	- working time - income - economic sector	- economic sector
Additional information asked (event-centred, every change tracked)		- temporary contract (yes/no) - employment contract	- temporary contract (yes/no) - employment contract - working time - income

The majority of the life history module consisted of interleafed looping questions (e.g. Eckman and Kreuter 2018) asking, for example, for the first employment in the reference period, then for details on this employment, and then for the next employment. For an independent random sample of the respondents, the looping questions were preceded by a life history calendar. The information from this matrix-like graphical calendar was used for routing the respondents into the relevant looping questions afterwards. This experiment is not part of

the presented article (see Lang and Carstensen 2022 for further details). Beyond employment episodes, structured and reported by the respondents, we asked for changes in the 'criteria splitting episodes into sub-episodes' (see Table 1), initialising further loops. This means, for instance, that an employment episode reported from January 2019 to December 2019 was split into two sub-episodes if the respondent indicated that they changed the work location in June 2019. All further information was then recorded for the resulting two sub-episodes. This and the sort of information asked were identical in the minimum, medium, and maximum information condition. To vary the response burden, the resolution of the requested information increased: we asked the respondents to indicate all changes within the loops for parts of the information (information is listed in the last row of Table 1, instead of the second row).

We decided to make the survey completely responsive (instead of a mobile-first design), meaning that the layout of the survey adapted to the devices of the respondents. For example, the lines were individually wrapped according to the available screen width and matrix questions switched to a carousel layout.

A total of 2,016 people responded to the questionnaire. Since we were also interested in break-offs in the life history module, we excluded only partials that dropped out before the life history module. Excluding such cases and those of respondents who did not consent to the usage of paradata left 1,518 cases to be analysed.

Our analyses encompass seven dependent variables. Since some models include respondents who dropped out of the questionnaire after they began the life history module, case numbers differ between the models. The panel consent could only be asked for a very small subset of the sample: those who were not already part of the access panel and did not terminate the survey earlier. In checking if the randomisation was successful, Tables 2 and 3 show descriptive statistics separately by the groups of both treatments.

Whether the composition of these observed properties differs by treatment group (p-value) was determined by OLS regression for metric and average marginal effects of logistic regressions for categorical variables. Surprisingly, the share differs between the groups of the detail treatment for one of the categories of current activities. We assume that this is a coincidence, since the current activity was asked at the very beginning of the questionnaire and respondents were only excluded when they broke off before the life history module, where the treatments first affected the response process. By controlling for these properties in our models, we compensate for this random error.

Table 4 presents an overview of the same compositional characteristics by type of device. Since this could not be randomised, differences in the composition by device are expected. Overall, there are no counterintuitive differences in this sense. There are two significant differences in the composition of these groups by subsample as well.

2.2. *Measurements*

This section documents how we operationalise each dependent and independent variable. In the independent part, all models include dummy variables for the treatment conditions and for mobile devices (1 = mobile device, 0 = other device), as well as interaction terms between the treatment conditions and the device variable. The treatment conditions were independently and randomly assigned, and the device was measured by coding the user agent string from our paradata. The user agent string was searched for "mobi", resulting in smartphones and tablets

Table 2. Composition by reference period treatment.

Variable	Two years				Five years				Difference mean two years versus five years
	Obs	Min/ Max	Mean	Std. Dev.	Obs	Min/ Max	Mean	Std. Dev.	P-Value
Mobile device	755	0/1	0.44	0.50	763	0/1	0.40	0.49	0.170
Current activity									
Student	755	0/1	0.49	0.50	763	0/1	0.51	0.50	0.607
Vocational training	755	0/1	0.06	0.24	763	0/1	0.04	0.20	0.056
Employed	755	0/1	0.43	0.50	763	0/1	0.44	0.50	0.774
PhD student	755	0/1	0.08	0.28	763	0/1	0.09	0.28	0.904
Other	755	0/1	0.15	0.36	763	0/1	0.19	0.39	0.067
Sample	755				763				
Access panel: (former) students		0/1	0.81	0.40		0/1	0.83	0.37	0.152
Vocational schools: school leavers		0/1	0.07	0.26		0/1	0.07	0.25	0.719
University: international students		0/1	0.07	0.26		0/1	0.07	0.25	0.574
University: non-traditional students		0/1	0.05	0.22		0/1	0.03	0.18	0.115
Gender: female	564	0/1	0.57	0.50	580	0/1	0.56	0.50	0.676
Age	467	17-60	28.34	5.57	438	18-56	28.3	5.72	0.908

Table 3. Composition by level of detail treatment.

Variable	Minimum detail				Medium detail				Difference mean med. to min.	Maximum detail				Difference mean max. to min.
	Obs	Min/ Max	Mean	Std. Dev.	Obs	Min/ Max	Mean	Std. Dev.	P-Value	Obs	Min/ Max	Mean	Std. Dev.	P-Value
Mobile device	501	0/1	0.44	0.50	508	0/1	0.41	0.49	0.252	509	0/1	0.41	0.49	0.296
Current activity														
Student	501	0/1	0.51	0.50	508	0/1	0.47	0.50	0.245	509	0/1	0.52	0.50	0.806
Vocational training	501	0/1	0.07	0.25	508	0/1	0.04	0.20	0.088	509	0/1	0.04	0.20	0.087
Employed	501	0/1	0.40	0.49	508	0/1	0.43	0.49	0.401	509	0/1	0.47	0.50	0.024
PhD student	501	0/1	0.09	0.28	508	0/1	0.1	0.31	0.316	509	0/1	0.07	0.25	0.255
Other	501	0/1	0.17	0.38	508	0/1	0.19	0.39	0.522	509	0/1	0.15	0.36	0.426
Sample	501				508					509				
Aaccess panel: (former) students		0/1	0.81	0.39		0/1	0.81	0.39	0.852		0/1	0.83	0.37	0.347
Vocational schools: school leavers		0/1	0.08	0.27		0/1	0.07	0.26	0.673		0/1	0.06	0.24	0.234
University: international students		0/1	0.08	0.27		0/1	0.06	0.24	0.293		0/1	0.07	0.25	0.58
University: non-traditional students		0/1	0.03	0.18		0/1	0.05	0.22	0.134		0/1	0.04	0.19	0.65
Gender: female	391	0/1	0.56	0.50	375	0/1	0.58	0.49	0.504	378	0/1	0.54	0.50	0.673
Age	320	17-51	28.09	5.43	287	18-60	28.18	5.88	0.838	298	18-51	28.71	5.62	0.172

Table 4. Composition by type of device.

Variable	Non-mobile				Mobile				Difference mean mobile versus non-mobile
	Obs	Min/ Max	Mean	Std. Dev.	Obs	Min/ Max	Mean	Std. Dev.	P-Value
Current activity									
Student	883	0/1	0.50	0.50	635	0/1	0.50	0.50	0.842
Vocational training	883	0/1	0.04	0.18	635	0/1	0.07	0.26	0.001
Employed	883	0/1	0.42	0.49	635	0/1	0.44	0.50	0.462
PhD student	883	0/1	0.11	0.31	635	0/1	0.05	0.22	0.000
Other	883	0/1	0.18	0.38	635	0/1	0.16	0.36	0.261
Sample	883				635				
Access panel: (former) students		0/1	0.85	0.36		0/1	0.78	0.42	0.000
Vocational schools: school leavers		0/1	0.04	0.20		0/1	0.11	0.31	0.000
University: international students		0/1	0.07	0.26		0/1	0.06	0.24	0.422
University: non-traditional students		0/1	0.04	0.18		0/1	0.05	0.22	0.114
Gender: female	669	0/1	0.52	0.50	475	0/1	0.62	0.49	0.001
Age	519	18-60	28.85	5.50	386	17-51	27.61	5.76	0.001

coded as mobile devices. In addition, all models are controlled for another experiment, including the (independent, random) assignment of a life history calendar in advance of the life history module itself, which is used to filter respondents into relevant detail questions. This experiment is not part of the presented article, but we include the treatment conditions as control variable to account for possible confounding through an effect on response burden as well as our dependent variables. (For results on the use of a life history calendar in a self-administered mixed-device online survey, see Lang and Carstensen 2022). Beyond that, all models are controlled for current activity (dummy variables for studying, vocational training, employment, PhD, others); sample (dummy variables for the samples described above); gender (1 = female, 0 = male); and age (in years).

As outlined in Subsection 1.2, the first analyses are devoted to testing hypotheses H1a and H1b and the question of whether the treatments lead to higher values on rather direct measures of response burden. Response time is measured in minutes and calculated from the paradata of our web survey as the sum of minutes spent on each page of the questionnaire. This time was corrected for interruptions, which are defined as a duration of more than five minutes on one page. The corrected response time is calculated as the total time spent in the questionnaire minus the time exceeding five minutes on each page. An exception was made for pages including a life history calendar (LHC). In this case, the threshold was set to ten minutes. For the other dimension, the subjective difficulty of the questionnaire, respondents were asked on a seven-point endpoint labelled "Likert scale" (0 = very easy to 6 = very difficult): "How difficult was it for you to provide the information on your background that we asked you about in this survey?"

For the second group of hypotheses on response behaviour, we investigate the probability that a respondent would break off the survey, and the probability that the participant would provide panel consent for further investigations. "Breaking off" is defined as terminating the questionnaire without resuming it before the end of our field period (0 = completion, 1 = break-off). The panel consent was asked for at the end of the questionnaire, on the same page as the contact details for the lottery (0 = no, 1 = yes). Unfortunately, respondents from subsample (1) could not be asked for their consent, as they were already part of the access panel.

For the third group of hypotheses, we investigate the share of months with no information, the number of episodes, and the number of INR as indicators for data quality. For the share of months without information, we use data from the life history calendar (if applicable) and the life history module. We count every month as being "with no information" if there is no episode that includes that month, regardless of the type of episode. This count serves as the nominator, while the denominator is the maximum number of months for which one respondent could provide information. The number of episodes is the simple count of all episodes a respondent provided, regardless of the type of episode; and the number of INR is a simple count of all items where answers are missing. Besides the variables mentioned above, these models include dummy variables that indicate the point of break-off in the life history module (if applicable), or the logarithmised total number of items each respondent received.

2.3. Analysis Strategy

To test our hypotheses, we run several regression analyses, each adapted to a single hypothesis. We model the reference period and the information treatment separately for

every dependent variable, as we are interested in the single effects of the treatments. Otherwise, we would have to apply the same sample restrictions on the models for the reference period as on the models for the information treatment. Due to the fully-crossed design and independent randomisation, this is not a problem.

We analyse response time, subjective difficulty, dropout, panel consent, number of episodes, share of months without an episode, and the number of INR using poisson, ordinary least squares (OLS), logistic, fractional, double hurdle, and negative binomial regression models. All models are controlled for current activity, sample, gender, and age. Those control variables are necessary because the treatment variables are randomised but the use of mobile versus non-mobile devices to answer the survey could not be allocated at random and former research reported that younger and female participants are more likely to use mobile devices (see De Bruijne and Wijnant 2014; Gummer and Roßmann 2015; Gummer et al. 2019; Sommer et al. 2017). Since we want to check whether the effects of our treatments differ between mobile and non-mobile devices, we thereby try to control for selection bias deriving from device-specific preferences. However, we cannot state with certainty that no selection bias remained after controlling for these variables.

For the models on response burden measured by interview duration and subjective difficulty, we apply poisson regression to implement a log-level-model, and OLS for a linear regression model. To test our hypotheses on response behaviour measured by dropout and panel consent, we run logistic regression models. For the models on data quality, we use more diverse estimators. Moreover, to model the share of months without information, we use a fractional response model with a logistic link function (Papke and Wooldridge 1996). This sort of model is appropriate for dependent variables with values ranging from 0 to 1, such as fractions or ratios.

By inserting dummy variables for the point of break-off within the life history module, we can include cases which answered only part of the module but do not confound our estimate with another mechanism. To model the number of episodes and the number of INR, we adopt Cragg's method of double hurdle regression (Cragg 1971; Burke 2009). Differently from the tobit model, this approach allows one to model the probability of a positive value and actual value (given that it is positive) separately, accounting for the possibility that they are caused by different mechanisms. In the models analysing the level of detail treatment, we only include cases where at least one episode was reported in the employment module; for this reason, we use negative binomial regression here. The models dealing with these two outcomes are controlled for the point of break-off dummies, as well as the logarithmised total number of items each respondent received. The total number of items controls for the sheer possibility of having a higher number of INR with an increasing number of questions being asked, and the logarithmic form allows for modelling a diminishing marginal utility. Besides theoretical reasoning, an empirical comparison with linear and quadratic forms also revealed best model fit with the logarithmic form.

Beyond the case selection mentioned above, we apply *multiple imputation by chained equations* (Van Buuren 1999, 2012; White et al. 2011) and the *multiple imputation then deletion* approach of Von Hippel (2007). We did $m = 50$ imputations with $i = 100$ iterations each, using propensity mean matching with the five nearest neighbours (Morris

et al. 2014; Vink et al. 2014; White et al. 2011) to predict plausible values. Additional auxiliary variables used in the imputation but not included in any of the models below are: day of the week and time of the first and last contact as well as their interactions, height and width of the device used, number of children, indicator for a general higher education entrance qualification and the year of obtaining the higher education entrance qualification.

For most variables, there is not a problem of missing values. Nevertheless, two variables have relevant shares of missing information, and we have to control for them due to the selectivity of mobile versus non-mobile users. These are gender and age, with 24% and 25% of information missing, respectively. Almost all the cases with missing information stem from subsample (1) and go back to the master data of the access panel. This variable was coded incorrectly for one of two samples of origin in the access panel. Therefore, we code the cases with implausibly high values on age (above 65 years) as 'missing', and impute plausible values for those cases, as well. Although the nature of the coding error is rather obvious from the distribution of the cases with an age above 65, we used mutiple imputation instead of simply recoding these cases, as it takes uncertainty into account.

3. Results

We estimate two models for each dependent variable, one for the reference period treatment and one for the detail treatment. Since we use double hurdle regression with two parts to test hypotheses H3c, there are 17 models in total. Table 5 shows the average marginal effects of those 17 different models.

All models contain an interaction term between the respective treatment effect and the dummy, specifying whether a mobile or a non-mobile device was used. Taking this interaction into account, the average marginal effects of the treatments are displayed in Figure 2. Additionally, we test the difference between the treatment effects for mobile and non-mobile devices, and the resulting *p*-values can be found in Table 6.

In what follows, we discuss the results for every dependent variable in detail. We first refer to the effects of the reference period treatment, then to the effects of the detail treatment, and finally move to the more differentiated view by type of device. We start by examining the first two hypotheses regarding the impact of our experiment on two dimensions of response burden (H1a and H1b).

3.1. Response Burden (H1a, H1b)

Models 1 and 2 in Table 5 address the time needed to complete the questionnaire (H1a). We estimate those by applying poisson regressions with robust standard errors. The results show that both treatments lead to a significantly increased duration in processing the questionnaire. The effect of the *reference period treatment* indicates an increase of the response duration in the five-year condition, compared to the two-year condition (model 1). The medium and maximum detail conditions of the *detail treatment* lead to an increase, as well, compared to the minimum detail condition (model 2). Looking at the *interactions between type of device and the treatments* in Figure 2, we see that, somewhat surprisingly, response duration is affected only by the treatments in participants who used non-mobile devices. However, the differences in treatment effects between mobile and non-mobile

Table 5. Regression models

	Response time		Difficulty		Break-off		Consent		Number of episodes			Share of months w/o episode		Item nonresponse			
	Poisson regression		OLS		Logistic regression		Logistic regression		Double hurdle linear regression		Negative binomial regression	Fractional response logistic regression		Double hurdle exponential regression			
	Treatment 1	Treatment 2	Treatment 1	Treatment 2	Treatment 1	Treatment 2	Treatment 1	Treatment 2	Treatment 1 Conditional response	Treatment 1 Pr(Y > 0)	Treatment 2	Treatment 1	Treatment 2	Treatment 1 Conditional response	Treatment 1 Pr(Y > 0)	Treatment 2 Conditional response	Treatment 2 Pr(Y > 0)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
Reference period: five years	0.128***		0.167		0.078***		0.041		1.286***	0.005		0.028*		0.039	-0.017		
(ref.: two years)	(4.65)		(1.71)		(3.34)		(0.58)		(11.59)	(0.36)		(2.00)		(0.27)	(-0.70)		
detail: medium		0.077*		-0.046		0.061*		0.277**			-0.140		-0.015			0.106	0.060
(ref.: minimum)		(2.46)		(-0.34)		(2.47)		(2.66)			(-0.77)		(-1.44)			(0.45)	(1.52)
detail: maximum		0.104**		-0.141		0.054*		-0.040			0.095		0.001			1.047***	0.226***
(ref.: minimum)		(3.05)		(-1.04)		(2.18)		(-0.42)			-0.52		(0.09)			(4.05)	(5.64)
mobile device	0.050	0.130***	0.163	0.071	0.124***	0.047	-0.018	0.183	-0.206	-0.025	-0.190	0.063***	0.016	0.099	0.052*	0.239	0.053
(ref.: non-mobile device)	(1.73)	(4.34)	(1.60)	(0.60)	(5.04)	(1.96)	(-0.23)	(1.53)	(-1.80)	(-1.78)	(-1.20)	(4.30)	(1.59)	(0.67)	(2.04)	(1.13)	(1.58)
Observations	1,518	892	1,053	789	1,518	892	163	82	1,518	1,518	892	1,518	892	1,518	1,518	892	892
additionally controlled for *break-off module*	x	x							x	x	x	x	x	x	x	x	x
ln no. of items seen in total														x	x	x	x
no. of job episodes		x															

Notes: Average Marginal Effects; t statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; all models controlled for: treatment calendar vs. q-list; current activity (studying, in vocational training, employment, PhD-Studies, internship, stay abroad, parental leave, unemployed, other); gender; age

Table 6. Test of interaction effects of independent variable with type of device.

	Response Time	Difficulty	Break-off	Consent	Number of spells	Pr(No. of spells > 0)	Share of Months w/o Spell	Item Nonresponse	Pr(No. of INR > 0)
Reference period: five years (ref.: two years)	0.1614	0.0346	0.0532	0.1689	0.0076	0.3038	0.6111	0.0044	0.0042
detail: medium (ref.: minimum)	0.1603	0.0544	0.0749	0.5223	0.8271	-	0.2833	0.8968	0.0318
detail: maximum (ref.: minimum)	0.0185	0.0639	0.0910	0.0191	0.2382	-	0.5356	0.5517	0.2067

Note: P-Values, Hypothesis test on difference between treatment conditions non-mobile vs. mobile

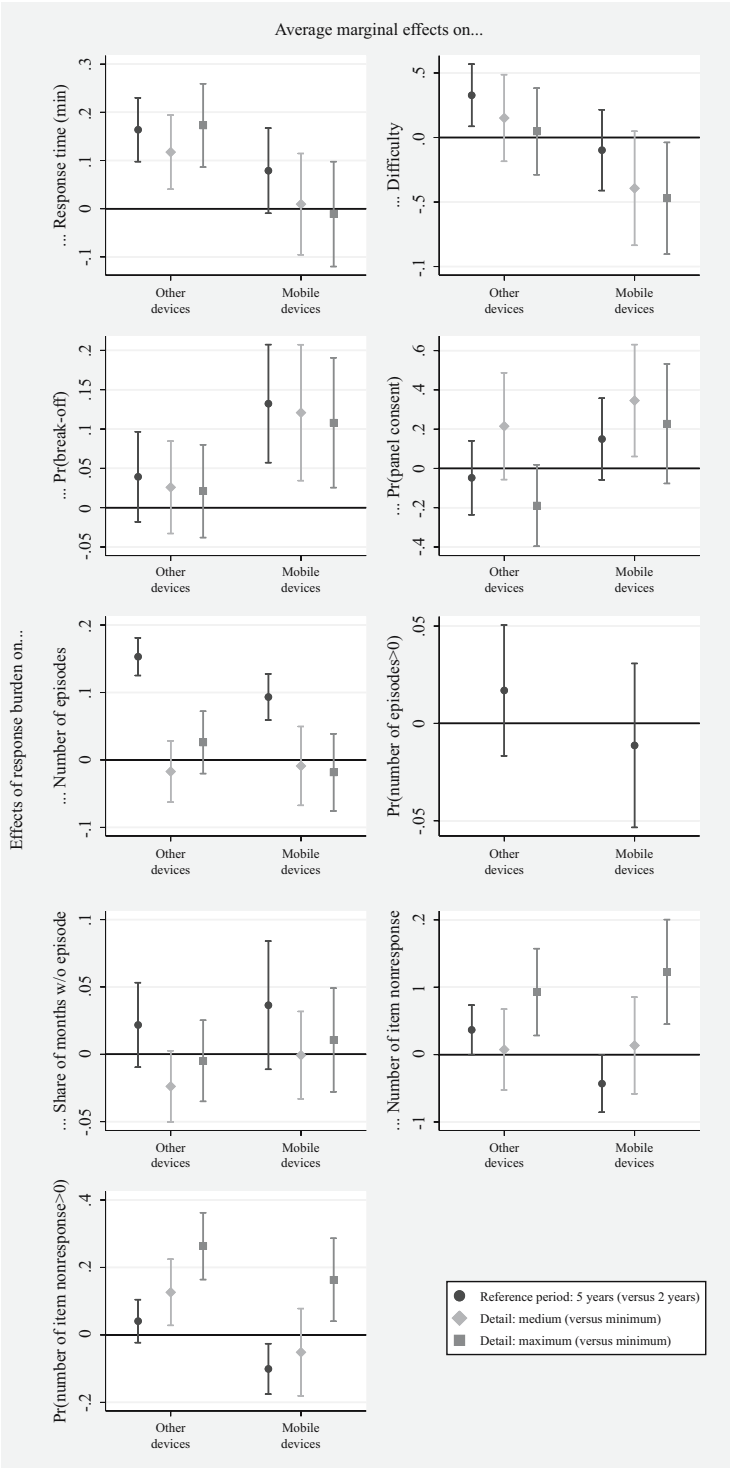


Fig. 2. Average marginal effect plots of interactions with the type of device.

users are significant only for the maximum condition of the detail treatment. One reason for this might be a higher tendency among mobile users to satisfice. But although we controlled for age, gender, and current activity, it is still possible for users of mobile and non-mobile devices to differ from each other in a way that affects their life history, thus leading to different interview durations.

Models 3 and 4 deal with the perceived difficulty of the questionnaire (H1b). The effects of the *reference period* (model 3) and *detail treatment* (model 4) are not significant. Looking at the *interactions*, there is a significant positive effect of the reference period treatment on perceived difficulty only for non-mobile users, and a negative effect of the maximum detail condition for mobile users. The treatment effect of the reference period differs significantly between mobile and non-mobile users, while differences in the effects of the detail treatment by device are only marginally significant.

3.2. Response Behaviour (H2a, H2b)

Here, we discuss the results concerning our hypotheses on response behaviour. In models 5 and 6, we consider the probability that respondents break off the survey (H2a). Increasing the *reference period* (model 5) leads to an increase of almost eight percentage points in this probability. The effects of the *medium and maximum detail conditions* (model 6) are quite similar: each one shows an increase of around five to six percentage points over the minimum detail condition. Looking again at the *interaction* between type of device and treatment effects, it appears that the effect is driven by the mobile users. Referring to the previous results on response burden, the higher likelihood that mobile users break off the survey could explain why there is no increase in response time in this group. However, this does not explain the difference between devices concerning perceived difficulty since this question was asked at the very end of the survey, and it was not seen by respondents who broke off.

Models 7 and 8 address the probability of consenting to future surveys (H2b). The consent question was administered only to respondents who were not already part of our access panel. As a result, case numbers for these two models are rather low. While we do not find a significant effect of the *reference period treatment* on the probability to give consent for future surveys (model 7), we do find a positive effect on consenting to future surveys in the medium condition of the *detail treatment* (model 8), stemming from mobile users. At the same time, looking at the *interactions*, the difference in treatment effects on mobile versus non-mobile users is not significant, given the low case numbers.

3.3. Data Quality (H3a, H3b, H3c)

We now turn to our hypotheses concerning data quality. Models 9–11 address hypothesis H3a, concerning the number of reported episodes. A higher number of episodes is regarded as positively connected to data quality. We estimated the model for the reference period treatment using a double hurdle linear regression, due to excess zeros in the data. Model 9 represents the part of the model conditional on $y_i > 0$, while model 10 represents the probability of $y_i > 0$.

Both parts of the models use the same covariates. It is not surprising that increasing the *reference period* (model 9) leads to an increase in the number of episodes reported by the respondents, since a longer reference period should almost certainly contain more single

episodes than a shorter one. The probability of reporting zero episodes is not affected by the reference period treatment. The *level of detail* (model 10) has no effect on the number of episodes.

Looking at the *interactions with the type of device*, the size of the effect of the reference period treatment differs significantly between mobile and non-mobile users. This may be due to selection, but also to a stronger sensitivity on the part of mobile users to the increased cognitive burden posed by the five-year condition relative to the two-year condition.

We use the share of months in which no episode of any type is reported as another measure of data quality. Hypothesis H3b states that an increase in response burden should lead to an increase in gaps in the reported life history. The results show a positive effect of the *reference period treatment* on the share of months for which no activity was reported (model 12). The *detail treatment*, by contrast, shows no effect (model 13). Turning to the *interactions*, the confidence intervals of the separate effects of the reference period by type of device include zero, due to decreased power. Nevertheless, the point estimate is positive for both groups. For the detail treatment, we do not find significant differences in the treatment effects between devices.

Finally, we test the hypothesis that an increase in response burden leads to an increase in item nonresponse (H3c). Again, due to excess zeros, we use double hurdle regression with the same covariates to estimate both parts of the model. In addition, we include the logarithmised number of items seen in total, as this varies between participants due to heterogeneous life courses.

The *reference period treatment* (models 14 and 15) demonstrates no effect on the number of item nonresponse. Conversely, the *level of detail* has a positive effect on item nonresponse, although only for the maximum detail condition compared to the minimum level of detail. The effect is present both for the conditional part (model 16) and the selection part (model 17) of $\Pr(y_i > 0)$.

Considering the *interactions* of the reference period treatment, significant differences between devices is evident for both parts of the model. A reference period of five years compared to two years leads to a higher amount of item nonresponse (conditional part) for non-mobile users (with a P-value of 0.051, this effect is just out of range of the 95%-confidence level), and a lower one for mobile users (P-value: 0.047). The probability that item nonresponse occurs at all (selection part) is found to be lower in the five-year condition than in the two-year condition, but only for mobile users.

The interactions for the effect of the detail treatment on the amount of item nonresponse (conditional part) show that it is independent of the type of device that was used. For the probability that item nonresponse occurs at all (selection part), there is an additional positive effect of the medium detail condition relative to the minimum detail condition, but it is present only for non-mobile users. This difference between devices is significant. At the same time, the positive effect of the maximum detail condition does not differ significantly between devices.

3.4. Summary of the Findings

We conclude that the *reference period treatment* directly affects response burden, measured as the duration of processing the questionnaire (H1a), and also affects the

subjective evaluation of difficulty (H1b). Surprisingly, the *detail treatment* only affects response time, and not subjective difficulty.

Distinguishing between users of mobile and non-mobile devices yields significant differences for these two hypotheses. While non-mobile users mostly demonstrate the expected effects, the same is not true for mobile users. Furthermore, mobile users show no significant increases in response time or self-assessed difficulty, as opposed to an unexpected negative effect of the maximum detail treatment on the latter. This indicates that either mobile users' reactions to our treatments are different or that the composition of this group differ from that of non-mobile users.

The next two hypotheses pertain to the participants' response behaviour. H2a, which is related to the probability of a break-off, could be confirmed for both treatments. As opposed to H1a and H1b, however, significant effects are present only among mobile users. This might be due to the fact that the users of mobile and non-mobile devices employ different strategies to deal with a higher response burden, with more mobile users breaking off the questionnaire. However, this picture might also be driven by selection into devices, despite our endeavours to account for this.

Despite the very low case numbers, panel consent (H2b) shows some noteworthy effects. Among mobile users and in the average marginal effects without differentiation, there is an increased probability of consent for respondents in the medium detail treatment relative to the minimum detail condition. This is against our expectations in two ways: (1) we assumed an incremental increase in response burden over our detail treatments and (2) we expected a negative effect of a higher response burden. Unfortunately, we do not have a plausible, better explanation for this result than chance. However, the extremely low case numbers in this model suggest that its interpretation should not be overstressed.

The last three hypotheses deal with data quality. Hypothesis 3a is directed only towards the detail treatments, which show no effects. Increasing the reference period should lead to an increased number of episodes even without taking the data quality into account. Nevertheless, the difference in effects for users of mobile and non-mobile devices is remarkable. The number of episodes reported increased significantly less in mobile users, and this could not be accounted for by controlling for age, gender, or current activity. One explanation for this finding might be a higher tendency of underreporting in this group.

The second data-quality hypothesis (H3b) concerns the number of gaps in the reported episodes. There is evidence that increasing the reference period increases the share of months without information, while the detail treatment shows no significant effect on this outcome. Moreover, the analyses of H3c yield partially counterintuitive results. The detail treatment shows expected effects for both mobile and non-mobile users, in both parts of the model; but alternating the reference period has only a small and marginally significant effect on non-mobile users modelling the number of nonresponses conditional on $y_i > 0$, while having a negative effect on mobile users in both parts of the model. One explanation for this might be that respondents using mobile devices show types of satisficing that do not increase item nonresponse but do yield a lower number of reported episodes (underreporting). Since we control for the total number of items seen, this could result in a decrease in item nonresponse. This explanation would also be concurrent with the smaller increase in the number of reported episodes from models 9 and 10.

4. Discussion

This article investigates the question of how response burden variations in terms of the amount and detail of information requested on life history affect different outcomes in a self-administered, mixed-device web survey of a highly educated population. We find several effects which stress the importance of managing response burden in self-administered online surveys.

Our results indicate that, as we expected, our treatments clearly increase response burden measured as response time and break-off. Notably, the differences in results between users of mobile and non-mobile devices point to different experiences of and reactions to increased response burden. Mobile participants show no effect on response time but do show an increase in break-offs. Conversely, non-mobile participants show no effect on break-offs but do show an increase in response time.

The results on subjective difficulty and panel consent are less straightforward. In general, while the effects are in the expected direction among non-mobile users, mobile users show unexpected behaviours. However, it should be kept in mind that the analyses on panel consent are based only on a very small number of cases, and only on subsamples (3) and (4). Participants from subsample (1) could not be asked once again about their panel consent, as they came from a probability-based access panel.

Concerning our data quality measures, we find effects on all three outcomes, but not for every treatment. Results on the number of episodes are consistent with the rest of our findings. The number of reported episodes is expected to be higher when the reference period is increased from two years to five years. However, this increase has a smaller effect on respondents using mobile devices. This might be an indication of specific satisficing strategies used by this group. At the same time, this group exhibits a negative effect of prolonged reference period on item nonresponse.

Both findings could be explained by a higher tendency towards underreporting in the group of mobile users, thus leaving less opportunity for nonresponse. As before, although we aimed to account for selectivity in mobile and non-mobile usage, we cannot rule out that this effect could be traced back to a different composition of these groups' life histories.

The detail level treatment does not affect the number of episodes reported. Regarding item nonresponse, this treatment shows expected results. Increasing the complexity of the survey by raising the level of detail does affect item nonresponse, independently of the device used.

Two key findings emerge from our study. First, varying the amount and complexity of information collected in a mixed-device web survey has different effects on data quality and response behaviour. Increasing the amount of requested information apparently leads to satisficing in mobile users who report a lower number of episodes, while item nonresponse is mainly influenced by the complexity of the questionnaire. Second, mobile, and non-mobile users seem to experience and react differently to response burden. Mobile users tend to break off more easily, while non-mobile users spend more time answering the questionnaire and rating the questionnaire as more difficult. As stated above, mobile users in our sample are also more prone to satisficing.

Although our study relies on experimental variations, there are several limitations to our findings. The possibility of selection bias concerning differences between mobile and non-

mobile users forbids us to interpret differences in the effects as causal. We try to account for all relevant confounders, but we cannot rule out that some of our findings could be driven by selective usage of devices. Furthermore, our experimental treatments operationalising the amount and complexity of information are not necessarily transferable to other surveys, due to the specific nature of our study.

We constructed our questionnaire using an interleafed looping strategy on filter and follow-up questions. Since it is known that this leads to different respondent behaviour than a grouped approach (Eckman and Kreuter 2018), our results might have been different using the latter. Moreover, our sample, comprising mostly participants of an access panel, is on average rather young and highly educated. There is evidence that response burden is perceived to be higher in mobile devices for older and lower-educated participants (Yan and Tourangeau 2008; Couper and Kreuter 2013). Hence, in different samples, the effects we found might be even more pronounced.

It is unclear whether our treatments would have yielded the same effects in a different population, in a survey containing different topics and being conducted by another organisation. Nevertheless, due to these conditions, it is reasonable to assume that we have found somewhat lower effects than would be expected in the general population. Although our study lacks objective measures of data quality, we are confident that we made reasonable presumptions about the expected effects on our outcomes.

For survey practice, our results have major implications. An increase in the scope and complexity of surveys should be considered carefully, given the effects on different outcomes we demonstrated. Furthermore, these effects might lead to a differential and complex bias by the type of device which respondents use to complete the survey, since our results pointed to different experiences and consequences of increased response burden along this divide. We recommend that future studies on this theme create techniques capable of eliminating these differences.

5. References

American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Available at: https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (accessed December 2020).

Antoun, C., and A. Cernat. 2020. "Factors Affecting Completion Times: A Comparative Analysis of Smartphone and PC Web Surveys." *Social Science Computer Review* 38 (4): 477–489. DOI: https://doi.org/10.1177/0894439318823703.

Antoun, C., M.P. Couper, and F.G. Conrad. 2017. "Effects of Mobile Versus PC Web on Survey Response Quality." *Public Opinion Quarterly* 81 (S1): 280–306. DOI: https://doi.org/10.1093/poq/nfw088.

Bradburn, N.M. 1979. "Respondent Burden." In *Health Survey Research Methods: Second Biennial Conference, Williamsburg*, edited by L. Reeder. Washington, D.C. U.S. Government Printing Office.

Burke, W.J. 2009. "Fitting and Interpreting Cragg's Tobit Alternative Using Stata." *Stata Journal* 9 (4): 584–592. Available at: https://www.stata-journal.com/article.html?article = st0179.

Burton, S., and E. Blair. 1991. "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys." *Public Opinion Quarterly* 55 (1): 50. DOI: https://doi.org/10.1086/269241.

Busemeyer, J.R., and J.T. Townsend. 1993. "Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment." *Psychological review* 100 (3): 432–459. DOI: https://doi.org/10.1037/0033-295X.100.3.432.

Buskirk, T.D., and C.H. Andrus. 2014. "Making Mobile Browser Surveys Smarter." *Field Methods* 26 (4): 322–342. DOI: https://doi.org/10.1177/1525822X14526146.

Clement, S.L., M. Kappelgaard Severin-Nielsen, and D. Shamsiri-Petersen. 2020. "Device Effects on Survey Response Quality. A Comparison of Smartphone, Tablet and PC Responses on a Cross Sectional Probability Sample." *Survey Methods: Insights from the Field*. DOI: https://doi.org/10.13094/SMIF-2020-00020.

Couper, M.P., and F. Kreuter. 2013. "Using Paradata to Explore Item Level Response Times in Surveys." *Journal of the Royal Statistical Society:* 176 (1): 271–286. DOI: https://doi.org/10.1111/j.1467-985X.2012.01041.x.

Couper, M.P., and G.J. Peterson. 2017. "Why Do Web Surveys Take Longer on Smartphones?" *Social Science Computer Review* 35 (3): 357–377. DOI: https://doi.org/10.1177/0894439316629932.

Cragg, J.G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica* 39 (5): 829. DOI: https://doi.org/10.2307/1909582.

Crawford, S,D., M.P. Couper, and M.J. Lamias. 2001. "Web Surveys." *Social Science Computer Review* 19 (2): 146–162. DOI: https://doi.org/10.1177/089443930101900202.

Daikeler, J., R.L. Bach, H. Silber, and S. Eckman. 2020. "Motivated Misreporting in Smartphone Surveys." *Social Science Computer Review*. DOI: https://doi.org/10.1177/0894439319900936.

De Bruijne, M, and M. Oudejans. 2015. "Online Surveys and the Burden of Mobile Responding." In *Survey Measurements: Techniques, Data Quality and Sources of Error*, edited by U. Engel: 130–145. Frankfurt, New York: Campus.

De Bruijne, M., and A. Wijnant. 2014. "Mobile Response in Web Panels." *Social Science Computer Review* 32 (6): 728–742. DOI: https://doi.org/10.1177/0894439314525918.

Drasch, K., and B. Matthes. 2013. "Improving Retrospective Life Course Data by Combining Modularized Self-Reports and Event History Calendars: Experiences from a Large Scale Survey." *Qual Quant* 47 (2): 817–838. DOI: https://doi.org/10.1007/s11135-011-9568-0.

Eckman, S., and F. Kreuter. 2018. "Misreporting to Looping Questions in Surveys: Recall, Motivation and Burden." *Survey Research Methods* 12 (1): 59–74. DOI: https://doi.org/10.18148/srm/2018.v12i1.7168.

Galesic, M. 2006. "Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22 (2): 313–328. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dropouts-on-the-web-effects-of-interest-and-burden-experienced-during-an-online-survey.pdf.

Galesic, M., and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Social Science Computer Review* 73 (2): 349–360. DOI: https://doi.org/10.1093/poq/nfp031.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2011. *Survey Methodology*. 2nd edition. Hoboken: John Wiley & Sons. Available at: http://gbv.eblib.com/patron/FullRecord.aspx?p = 819140.

Guin, T.D.-Le., R. Baker, J. Mechling, and E. Ruyle. 2012. "Myths and Realities of Respondent Engagement in Online Surveys." *International Journal of Market Research* 54 (5): 613–633. DOI: https://doi.org/10.2501/IJMR-54-5-613-633.

Gummer, T., F. Quoß, and J. Roßmann. 2019. "Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones?" *Social Science Computer Review* 37 (3): 371–384. DOI: https://doi.org/10.1177/0894439318766836.

Gummer, T., and J. Roßmann. 2015. "Explaining Interview Duration in Web Surveys." *Social Science Computer Review* 33 (2): 217–234. DOI: https://doi.org/10.1177/0894439314533479.

Hambleton, R.K., H. Swaminathan, and H.J. Rogers. 1991. *Fundamentals of Item Response Theroy*. California: Sage.

Hedlin, D., H. Lindkvist, H. Bäckström, and J. Erikson. 2008. "An Experiment on Perceived Survey Response Burden Among Businesses." *Journal of Official Statistics* 24 (2): 301–318. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-experiment-on-perceived-survey-response-burden-among-businesses.pdf.

Hoogendoorn, A.W., and D. Sikke. 1998. "Response Burden and Panel Attrition." *Journal of Official Statistics* 14 (2): 189–205. Available at: https://www.scb.se/contentassets/-ca21efb41fee47d293bbee5bf7be7fb3/response-burden-and-panel-attrition.pdf.

Kaper, E., and W. Saris. 1999. "Effects of Response Burden in a Continuous Survey." *Kwantitative Method* 61: 5–17.

Keusch, F., and T. Yan. 2017. "Web Versus Mobile Web." *Social Science Computer Review* 35 (6): 751–769. DOI: https://doi.org/10.1177/0894439316675566.

Krebs, D., and J.K. Höhne. 2020. "Exploring Scale Direction Effects and Response Behavior Across Pc and Smartphone Surveys." *Journal of Survey Statistics and Methodology* 9: 261. DOI: https://doi.org/10.1093/jssam/smz058.

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.

Lang, S., and J. Carstensen. 2022. "A life history calendar in a CAWI? Evidence from higher education research." In *Survey Methoden in der Hochschulforschung*, edited by G. Brandt and S. de Vogel: 121–152. Wiesbaden: Springer VS. DOI: https://doi.org/10.1007/978-3-658-36921-7_6.

Lee, H., S. Kim, M.P. Couper, and Y. Woo. 2019. "Experimental Comparison of PC Web, Smartphone Web, and Telephone Surveys in the New Technology Era." *Social Science Computer Review* 37 (2): 234–247. DOI: https://doi.org/10.1177/0894439318756867.

Lenzner, T. 2012. "Effects of Survey Question Comprehensibility on Response Quality." *Field Methods* 24 (4): 409–428. DOI: https://doi.org/10.1177/1525822X12448166.

Lugtig, P., and V. Toepoel. 2016. "The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey." *Social Science Computer Review* 34 (1): 78–94. DOI: https://doi.org/10.1177/0894439315574248.

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schütz. 2007. "Compensating for Low Topic Interest and Long Surveys." *Social Science Computer Review* 25 (3): 372–383. DOI: https://doi.org/10.1177/0894439307297606.

Mavletova, A. 2013. "Data Quality in PC and Mobile Web Surveys." *Social Science Computer Review* 31 (6): 725–743. DOI: https://doi.org/10.1177/0894439313485201.

Morris, T.P., I.R. White, and P. Royston. 2014. "Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws." *BMC medical research methodology* 14: 75. DOI: https://doi.org/10.1186/1471-2288-14-75.

Papke, L.E., and J.M. Wooldridge. 1996. "Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates." Journal of Applied Econometrics 11 (6): 619–632. DOI: https://doi.org/10.1002/(SICI)1099-1255(199611) 11: 6%3C619::AID-JAE418%3E3.0.CO;2-1.

Peytchev, A., and E. Peytcheva. 2017. "Reduction of Measurement Error Due to Survey Length: Evaluation of the Split Questionnaire Design Approach." *Survey Research Methods* 11 (4): 361–368. DOI: https://doi.org/10.18148/srm/2017.v11i4.7145.

Reise, S.P., A.T. Ainsworth, and M.G. Haviland. 2005. "Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research." *Current directions in psychological science* 14 (2): 95–101. DOI: https://doi.org/10.1111/j.0963-7214.2005.00342.x.

Schlosser, S., and A. Mays. 2018. "Mobile and Dirty." *Social Science Computer Review* 36 (2): 212–230. DOI: https://doi.org/10.1177/0894439317698437.

Sharp, L.M., and J. Frankel. 1983. "Respondent Burden: A Test of Some Common Assumptions." *Public Opinion Quarterly* 43. DOI: https://doi.org/10.1086/268765.

Sommer, J., B. Diedenhofen, and J. Musch. 2017. "Not to Be Considered Harmful." *Social Science Computer Review* 35 (3): 378–387. DOI: https://doi.org/10.1177/0894439316633452.

Toepoel, V., and P. Lugtig. 2014. "What Happens If You Offer a Mobile Option to Your Web Panel? Evidence from a Probability-Based Panel of Internet Users." *Social Science Computer Review* 32 (4): 544–560. DOI: https://doi.org/10.1177/0894439313510482.

Tourangeau, R. 2000. "Remembering What Happened: Memory Errors and Survey Reports." In *The Science of Self-Report: Implications for Research and Practice*, edited by A.A. Stone. Mahwah, N.J: Lawrence Erlbaum.

Tourangeau, R. 2003. "Cognitive Aspects of Survey Measurement and Mismeasurement." *International Journal of Public Opinion Research* 15. DOI: https://doi.org/10.1093/ijpor/15.1.3.

Tourangeau, R. 2018. "The Survey Response Process from a Cognitive Viewpoint." *Quality Assurance in Education* 26 (2): 169–181. DOI: https://doi.org/10.1108/QAE-06-2017-0034.

Van Buuren, S. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine* 18 (6): 681–694. DOI: https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6 < 681:AID-SIM71 > 3.0.CO;2-R.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton, Fla. CRC Press.

Vink, G., L.E. Frank, J. Pannekoek, and S. van Buuren. 2014. "Predictive Mean Matching Imputation of Semicontinuous Variables." *Statistica Neerlandica* 68 (1): 61–90. DOI: https://doi.org/10.1111/stan.12023.

Von Hippel, P.T. 2007. "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology* 37 (1): 83–117. DOI: https://doi.org/10.1111/j.1467-9531.2007.00180.x.

Warriner, G.K. 1991. "Accuracy of Self-Reports to the Burdensome Question: Survey Response and Nonresponse Error Trade-Offs." *Quality & Quantity* 25: 253–269. DOI: https://doi.org/10.1007/BF00167531.

White, I.R., P. Royston, and A.M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–399. DOI: https://doi.org/10.1002/sim.4067.

Yan, T., S. Fricker, and S. Tsai. 2020. "Response Burden: What Is It and What Predicts It?" In *Advances in Questionnaire Design, Development, Evaluation and Testing*, edited by P.C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G.B. Willis, and A. Wilmot: 193–212. Hoboken, NJ, USA: Wiley.

Yan, T., and R. Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22 (1): 51–68. DOI: https://doi.org/10.1002/acp.1331.

Zanon, C., C.S. Hutz, H. Yoo, and R.K. Hambleton. 2016. "An Application of Item Response Theory to Psychological Test Development." Psicologia: Reflexão e Crítica / Psychology: Research and Review 29 (1): 1. DOI: https://doi.org/10.1186/s41155-016-0040-x.

Received January 2021
Revised July 2021
Accepted December 2021