Investigating paradata for one of the largest surveys in Sweden

2024 Expert Meeting on Statistical Data Collection and Sources

Andreea Bolos
Viktor Dahl
Sofia Holsendahl

SCB

Table of contents

Abstract

There is a rapidly increasing and evolving landscape of online data collection, which enables a more time efficient and cost saving procedure. Statistics Sweden has adopted this digital opportunity to reach out to the citizens. However, a digital data collection comes with both perks and challenges and ensuring an effective process is paramount for delivering high quality data. Therefore, the aim of this paper was to study different typologies of web paradata and their role in improving data collection efficiency and ensuring a high data quality.

To conduct this study, we used data from one of Statistic's Sweden largest yearly survey, the Citizen's survey, commissioned by Sweden's municipalities, summing in around 210 000 citizens in the sample group. The survey encompassed a multi-mode approach, with both digital and paper data collection. The vast size of this survey provided us the opportunity to address paradata-oriented aspects. We addressed issues such as what devices respondents prefer and questionnaire navigation.

1. Introduction

1.1. Background and aim

There is a rapidly increasing and evolving landscape of online data collection, which enables a more time efficient and cost saving procedure. For an online survey, it is possible collect or access four types of data: (1) substantive data, (2) metadata, (3) paradata, and (4) auxiliary data (Callegaro, 2013). Substantive data represents the results of the questionnaire, the answers for each of the questions. Metadata represents the description of the project, the dates in which the survey was in the field, as well as other context information that could be relevant for the dataset, such as contact attempts for example. Paradata, also called process data, contains information about the respondent's survey-taking behavior such as contact process, device type, questionnaire navigation, and response time. Paradata are collected unobtrusively, meaning that participants are not deliberately providing this information. And finally, auxiliary data which are not gathered directly during the survey but attained from certain databases and can be used in the sampling process (Callegaro, 2013; Kunz, & Hadler, 2020). In our study, auxiliary data are comprised of socio-demographic variables that were used for sampling purposes and for statistical adjustments such as weighting. Auxiliary data were later merged with the substantive data, metadata, and paradata.

In this paper we use all four types of data to study the Statistics Sweden's citizens' survey. We consider this an exploratory analysis where we incorporate different typologies of data and explore their role in improving data collection and ensuring a high data quality. We aim to address issues such as what devices respondents use and response times. More specifically, our analysis addresses three types of paradata: contact information, device-type and questionnaire navigation paradata. In doing so, we address the following questions:

(1) Are specific socio-demographic characteristics of the study sample associated with longer or shorter response times? Based on previous research age is correlated with a lower cognitive ability and as well as working memory capacity (Couper & Kreuter, 2013). Moreover, education has been shown to affect response times, with longer response times for people with lower education (Couper & Kreuter, 2013).

(2) Are specific device types associated with longer or shorter response times?

(3) Do the citizens' attitudes towards their municipality have any effect on the response time? Previously, response times have been shown to be inversely correlated with the participants'

tendency to answer positively, which is known as acquiescence bias (Couper & Kreuter, 2013; Kuru & Pasek, 2016).

1.2. The Statistics Sweden’s citizens’ survey

The Statistics Sweden's citizens' survey is carried out to examine how residents feel about their municipality and the politicians and officials who works there. The survey has been conducted since 2005 and the sample population consists of residents of the municipalities aged 18 and older. The main survey consists of 41 questions divided in 14 different areas: (1) school and care services, (2) housing and neighborhood, (3) local employment and educational opportunities, (4) public service, (5) transport and communications, (6) library, cultural scene and meeting places, (7) sports, exercise and outdoor life, (8) maintenance of the public environment, (9) safety in society, (10) climate and environment work, (11) treatment, information and influence in the municipality, (12) equality and integration, (13) confidence and (14) general question about the municipality. Moreover, each municipality has the possibility of adding additional questions, based on their needs. The way the sampling has been done, the Statistics Sweden's citizen survey makes it possible to compare the opinions of different groups within each municipality, but also how a specific municipality stands relative to other municipalities and the entire country. Furthermore, Sweden has a total of 290 municipalities and in the year of 2023 a total of 162 municipalities participated in the citizens survey, which was the highest number of municipalities participating since 2005. That resulted in a sample of 208 000 citizens. The final response rate, pertaining both paper and digital answers, was 36 percent.

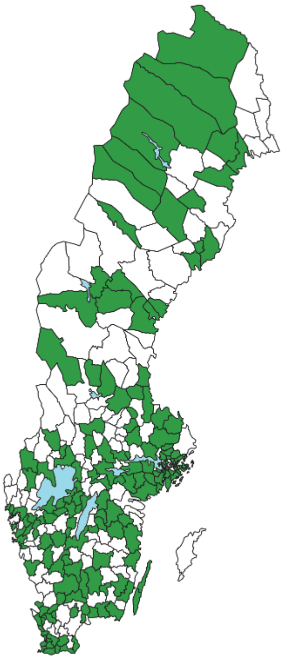


Figure 1. A map of Sweden's 290 municipalities. Green areas represent municipalities participating in the citizens' survey in 2023.

Since the beginning of the Statistics Sweden's citizen survey in 2005, the data collection strategy has always included paper questionnaires in two send-outs. In 2011, responders could for the first time answer the questionnaire online. Since 2021, respondents with a digital mailbox receive three send-outs to their digital mailbox and only one with a paper questionnaire. Those with a digital mailbox receive mail from authorities digitally instead of via postal mail.

2. Sample

Stratified random sampling was used in the survey and the sample size in 2023 was 208 000 with a response rate of 36 percent, including both paper and digital answers. Fifty-five percent answered the questionnaire online and 45 percent via paper.

The number of respondents and their response rate divided by different socio-demographics are shown in the table below. The table includes the responders who answered the questionnaire online. As our primary focus in this paper is response time, we decided to trunk the sample and to include only municipalities using the standard 41-question long questionnaire. By way of this procedure, we could ensure that all respondents in the analytical sample had received the same questionnaire.

Table 1. Socio-demographic characteristics of the study sample

	N	**%**
Gender		
• Male	14 316	47,4
• Female	15 862	52,6
Age		
• 18–29	4 131	13,7
• 30–49	8 141	27,0
• 50–64	11 078	36,7
• 64 or older	6 828	22,6
Income		
• Low	4 713	15,6
• Medium	15 172	50,3
• High	10 293	34,1
Education		
• Primary education	3 308	11,0
• Secondary education	12 867	42,6
• Higher education	14 003	46,4
Country of birth		
• Sweden	26 539	87,9
• Outside Sweden	3 639	12,1
Total	**30 178**	

3. Data preparation and analysis

Since this was an exploratory analysis, as a first step we decided to look at the final sample of completes without any data cleaning.

First, we looked at the descriptive statistics of the paradata by certain socio-demographic characteristics to get familiar with the data. We also had great help by the process data report for this survey developed by our colleagues Andersson & Stavås (2023).

Thereafter, as one potential way to clean our data, we looked at extreme values regarding respondents' response times. We discussed and concluded what could be an appropriate threshold for a minimum and maximum response time. In the end, however, we decided to include all responses – irrespective of if respondents answered unusually fast, or slow – in the further analysis of this paper.

As a final analysis we conducted an ordinary least squares regression (OLS) to better understand how one particular form of paradata – response time – related to some standard socio-demographic factors, the type of device used for answering the survey, and respondents' general attitude toward their municipality. This paper continues with the results of our analysis and thereafter ends with a short summary and some ideas for further discussion.

4. Results

This chapter is structured as follows; firstly, we present descriptive statistics of the different forms of web paradata, then we move forward to present the results of the OLS.

4.1. Contact information paradata

In total there were four contact attempts. As can be seen in the table below, the contact attempts were designed differently depending on age. The survey started on 08.24.2023 and ended on 07.11.2023.

Table 2. Contact attempts and completed answers

Send out	Respondents under 65 years old	Respondents 65 years old or older
1	Invitation with login to the web survey	Invitation with login to the web survey
2	Reminder	Reminder including a paper questionnaire
3	Reminder including a paper questionnaire	Reminder
4	Reminder	Reminder including a paper questionnaire

Amongst the 30 178 participants who answered the questionnaire online, most of them accessed the questionnaire only once. Moreover, we could see in the dataset that the time of the first log in and completion of the questionnaire varied between 0 and 70 days. However, the prevailing behavior amongst participants was that they accessed and completed the questionnaire within the same day.

4.2. Device type paradata

In total there were 383 (1 precent) respondents that changed operating system during their participation, 425 (1 precent) respondents changed web browser and 184 (0.6 precent) changed devices (e.g., from mobile phone to computer/laptop). Pertaining devices, the most frequent change took place between computer/laptop and mobile phone.

During the data collection, the most common device used was the mobile. While 65 percent of participants used a mobile phone to answer the survey, computer/laptop was used by 33 percent, and only 2 percent used a touchpad (see Figure 2). It is important to mention that participants could use several devices during the data collection. Therefore, there could be a many-to-one relationship between device and participants. The diagram below depicts the use of devices by participants.

Comparing by gender, mobile phone is still the most used device by both males and females, however more women (10 precent more relative to men) answered by mobile (see Figure 3). With regard to device type, a final comparison (Figure 4) shows that younger respondents seemed to prefer answering the survey via mobile. Mobile phone is the most used device type in all the three youngest age groups compared.

Figure 2. Use of devices total

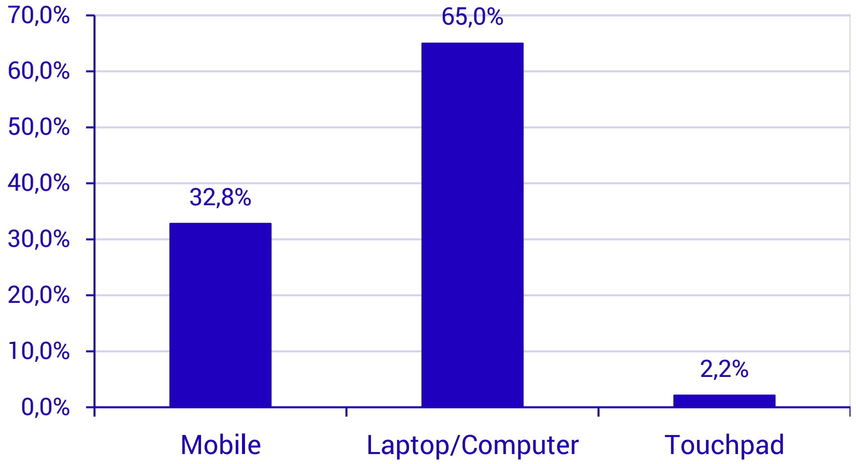


Figure 3. Use of devices per gender

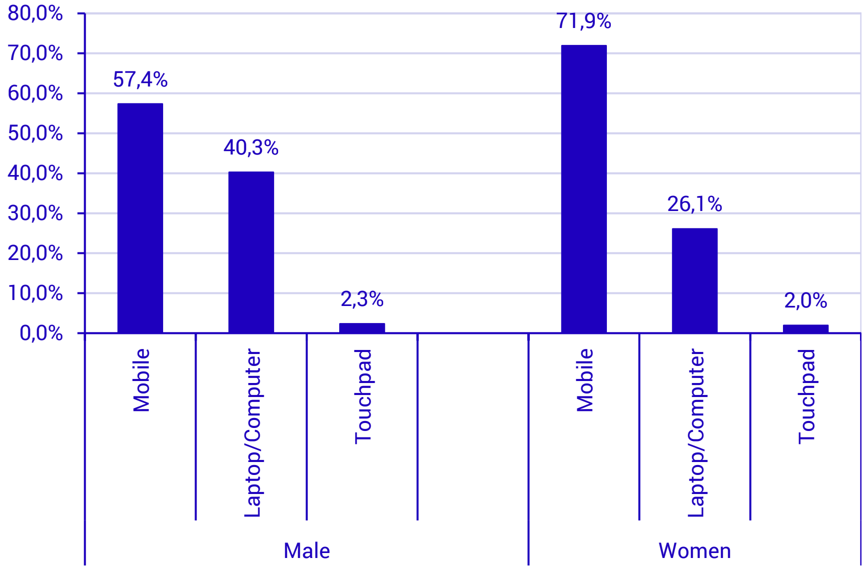
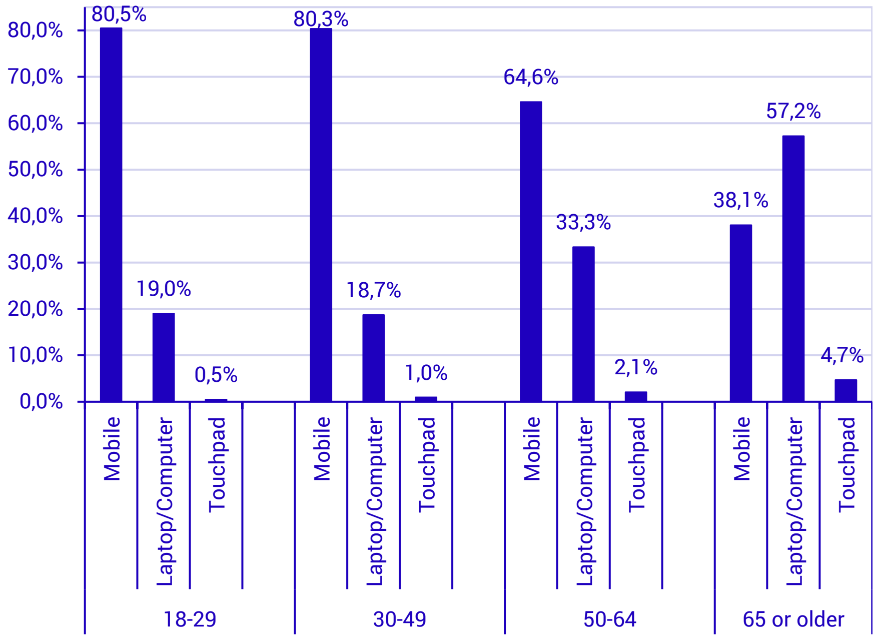


Figure 4. Use of devices per age group



4.3. Questionnaire navigation paradata

In the following, we report descriptive statistics of respondents' response time. The figures reported here were based on what we refer to as *Active time*. The measure of *Active time* refers to respondents' response time, in minutes. To be able to address the fact that respondents could login and logout as they pleased, we removed sections of time when no activity was recorded during instances of 20 minutes and above.

Regarding response times, Table 3 reports that both men and women tend to answer the survey in, on average, 20 minutes. Comparing over age, whereas younger respondents seem faster, respondents 65 years and older report slower response times on average.

Table 3. Active time

	N	Mean	Std Deviation	Median	Variance
Gender					
• Male	14 316	**20,7**	11,1	**17,8**	124,0
• Female	15 862	**20,1**	11,1	**17,1**	122,5
Age					
• 18–29	4 131	**19,1**	11,6	**15,7**	133,6
• 30–49	8 141	**19,2**	12,0	**15,9**	143,5
• 50–64	11 078	**19,9**	10,3	**17,2**	105,9
• 65 or older	6 828	**23,4**	10,4	**20,9**	109,1
Total	**30 178**	**20,4**	**11,1**	**14,6**	**123,3**

We also examined the extent to which respondents answered differently, in terms of response time, based on the device they used. As seen below (Table 4), respondents using mobile phones (mean = 19,3) answered the survey 3 minutes faster than respondents using laptops/computers (mean = 22,3), on average. The comparatively smaller group of respondents using touch pads answered in, on average, 21,8 minutes.

Table 4. Active time per device (only for those who did not change device)

	N	Mean	Std Deviation	Median	Variance
Laptop/computer	9848	**22,3**	12,1	**19,2**	146,2
Mobile	19 501	**19,3**	10,4	**16,6**	107,5
Touch Pad	645	**21,8**	10,2	**17,8**	103,1
Total	**29 994**	**20,3**	**11,1**	**17,43**	**122,1**

Extreme value analysis

As we mentioned before, response times have been found to be inversely correlated with the tendency to answer positively, irrespectively of the particular item (Couper & Kreuter, 2013). However, is it a fact that comparatively fast responses always can be characterized as careless? And are slow answers distracted, confused, or attentive (Read, Wolters, and Berinsky, 2021)? Additionally, perhaps unusually long or unusually short response times can be used as proxy indicators for measurement error, controlling for other factors that influence response times.

Table 5 below reports our comparison of three groups which were generated based on their response time. Both our prior information and this year's data showed that respondents tended to answer the survey in around 20 minutes. The survey questions vary, of course, but 20 minutes (20 * 60 = 1 200 seconds) equal to around 30 seconds per question. We decided that respondents answering faster than 5 minutes, which equals to 7 seconds per question was too fast. The three groups generated based on their response time were the following: (1) *Less than 5 minutes;* (2) *5 <= active time min <= 60*; (3)*Above 60 min.*

Although the total number of answers available to us are substantial, the choice to limit the group of fast respondents to those answering 5 minutes or shorter might be too strict. Table 5 shows that this group comprises only 29 individuals. A number as small as 29 individuals might be good from the perspective that the level of "speeders" seems affordable. However, further analyses including the *Less than 5 minutes*-group might be difficult. Moreover, table 5 also highlights that the share of respondents in the category – born outside of Sweden – are clearly overrepresented in the group of *Above 60 min*.

Finally, when comparing the groups generated from response time by levels of partial non-response, we found no differences comparing those who answered faster with others. However, faster respondents (i.e., those in the *Less than 5 minutes*-group) tended to mark *Don't know* as their answer to a higher degree than respondents in the other two groups.

Table 5. Extreme values analysis

	Less than 5 min ***n* (%)**	**5<=active time min<=60** ***n* (%)**	**Above 60 min** ***n* (%)**
Gender			
Male	20 (69,0)	14 137 (47,4)	159 (46,9)
Female	9 (31,0)	15 673 (52,6)	80 (53,1)
Age			
18–29	15 (51,7)	4 063 (13,6)	53 (15,6)
30–49	9 (31,0)	8 019 (26,9)	113 (33,3)
50–64	3 (10,3)	10 982 (36,8)	93 (27,4)
65 or older	2 (6,9)	6 746 (22,6)	80 (23,6)
Income			
Low	11 (37,9)	4 590 (15,4)	112 (33,0)
Medium	11 (37,9)	14 988 (50,3)	173 (51,0)
High	7 (24,1)	10 232 (34,3)	54 (15,9)
Education			
Primary	7 (24,1)	3 231 (10,8)	70 (20,7)

	Less than 5 min ***n* (%)**	**5<=active time min<=60** ***n* (%)**	**Above 60 min** ***n* (%)**
Secondary	11 (37,9)	12 726 (42,7)	130 (38,4)
Higher	11 (37,9)	13 853 (46,5)	139 (41,0)
Country of birth			
Sweden	19 (65,5)	26 318 (88,3)	202 (60,6)
Outside Sweden	10 (34,5)	3 492 (11,7)	137 (40,4)
N	**29**	**30 178**	**339**

4.5. Modelling response time

Our final endeavor for this paper was to model response time in order to examine the extent to which it could be explained by socio-demographic factors, device type, and respondents' general attitude to their municipality. We decided to run the above in an ordinary least squares (OLS) regression with response time as the dependent variable.

Socio-demographic factors and device type were already available to us, but we needed to create a general attitude toward the municipality measure. We generated respondents' municipality score from the last four questions in the survey: (1) How do you find your municipality as a place to live and work? (2) How do you find that your municipality runs its various operations? (3) Do you think the inhabitants of your municipality have the possibility of insight an influence over the municipality's decisions and operations? (4) Do you recommend that others, who do not live in the municipality, move here? For doing so we run first a factor analysis, which indicated that one score could be created, with one factor having the Eigen value of 2.2.

Table 6 contains the results from the OLS regression, where we had response time as dependent variable. Looking at the item level predictors we can see that relative to the last age group, 65 or older, the first two younger groups (18–29 and 30–49 years) answer significantly faster, almost five and four minutes faster. This is in accordance with previous research. Moreover, the education level does seem to influence the response time. Relative to a higher education level, a lower education level being associated slower response times, around a minute difference. When it comes to the device type, we removed the group that answer through touch pad, due to a small sample. Analyzing the difference between laptop and mobile, we can see that relative to mobile, participants that answer on a laptop were, on average, around two minutes slower. Finally, there were no differences between genders and the municipality score had no effect on respondents' response time.

Table 6. OLS

Parameter	Estimate	Standard Error	t Value	Pr>|t|
Intercept	25,45	0,39	64,95	***
Gender				
• Male	0,32	0,13	2,48	
• Female	Ref	Ref	Ref	Ref
Age				
• 18–29	-4,71	0,24	-19,18	***
• 30–49	-3,47	0,19	-18,07	***
• 50–64	-2,63	0,17	-14,89	
• 65 or older	Ref	Ref	Ref	Ref
Country of birth				
• Sweden	-5,33	0,19	-27,12	***
• Outside Sweden	Ref	Ref	Ref	Ref
Education				
• Primary	1,11	0,22	11,72	***
• Secondary	0,64	0,13	12,45	***
• Higher	Ref	Ref	Ref	Ref
Income				
• Low	2,72	0,23	11,72	***
• Medium	1,85	0,14	12,45	***
• High	Ref	Ref	Ref	Ref
Device				
• Laptop	1,95	0,142	13,77	***
• Mobile	Ref	Ref	Ref	Ref
Municipality score	-0,12	0,08	-1,56	
Model: p <0,0001				

NB: Comparisons significant at the 0.05 level are indicated my ***
Low R^2 value of 6 percent. We consider this to be ok, since we are mainly interested in the association between the predictors and response time and not in predicting the response time.

5. Discussion and conclusion

In this paper we use four types of data to study the Statistics Sweden's citizens' survey. We addressed our research questions with an exploratory design and incorporated different typologies of paradata in our attempts to answer whether (1) specific socio-demographic characteristics of the study sample are associated with longer or shorter response times; (2) specific device types are associated with longer or shorter response times; and (3) whether the citizens' attitudes towards their municipality have any effect on the response time.

With regard to socio-demographic characteristics, we found some indications that age, country of birth, income, and education all could be factors to investigate further. Moreover, it seems that respondents using mobile phones while answering the survey do so comparatively faster than those answering via computer/laptop. However, the idea that respondents' general attitude toward their municipality could be related to response time was not supported.

As mentioned, this paper had an explorative design, and our results are uncertain. Despite that, it is interesting to consider how age differences in response times matter. Is it that faster responses come with, on average, less attentiveness and, the opposite, that slower response means that respondents are observant? Or is it a curvilinear relation? One possible way to hinder too fast responses is to include so called screener questions. However, such attempts might instead result in greater respondent fatigue.

In this paper we used an aggregate measure of response time. This has limitations, of course. The aggregate measure of response time does not indicate whether specific questions are more difficult to answer or if specific sections of the questionnaire require more effort for certain subgroups. One way forward is therefore to combine aggregate measures of response time with efforts more oriented toward per-question analyses of response time. Another endeavor is to compare so called drop-outs with those who competed the survey. By way of such procedure, we could better understand peculiarities with our questionnaire and hopefully address matters which is difficult for respondents answering the survey.

Citations

1. Andersson G. & Stavås E. (2023). Tutorial for paradata reports, Statistics Sweden.
2. Kunz, T. & Hadler, P. (2020). Web Paradata in Survey Research. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS - Survey Guidelines).
3. Kurur, O. & Pasek, J. (2016). Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research, Computers in Human Behavior, Volume 57, Pages 82-92
4. Mick P. Couper & Frauke Kreuter (2013). Using paradata to explore item level response times in surveys, Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 176, No. 1, pp. 271-286
5. Read, B., Wolters, L., & Berinsky, A. (2021). Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys. Political Analysis, pp. 1-20.