Visualizing Survey Flow and Improving Data Collection Through Paradata

Gezim Seferi [1], Bengt Oscar Lagerstrøm [2]

[1] *Department of data collection, Statistics Norway (SSB)*

[2] *Department of data collection, Statistics Norway (SSB)*

Abstract

The transition from Computer-Assisted Telephone Interviewing (CATI) to Computer-Assisted Web Interviewing (CAWI) signifies a critical shift in survey methodology, driven by a need for cost efficiency and greater respondent autonomy. This study investigates the implications of this transition, focusing on "Nonresponse Error" within the Total Survey Error (TSE) model. Utilizing the Cross-border shopping survey conducted by Statistics Norway as a case study, this research explores the effectiveness of paradata in addressing challenges such as survey dropouts, data underreporting, and questionnaire bottlenecks that are prevalent in the CAWI format.

In the absence of an interviewer to guide respondents through the process, the design and clarity of questions in a CAWI environment are of utmost importance. Paradata, which includes detailed records of keystrokes, mouse movements, and timestamps, proves invaluable in identifying problematic elements of the survey. This type of data enables visualization tools such as the Sankey diagram to map out the journey of respondents through the survey, highlighting critical areas where dropouts are significant and where respondents are likely to backtrack or exit the survey prematurely.

The Cross-border shopping survey analysis pinpointed sections with repetitive questions about expenditures during multiple day trips abroad as frequent points of respondent dropout. The visualizations clearly demonstrated how the sequence and formulation of questions significantly impacted respondent engagement and response quality, indicating a complex relationship between questionnaire design and user interaction.

Further investigation into paradata revealed that respondents frequently changed their answers. This tendency to modify responses was analyzed by comparing timestamps and answers, where a notable proportion of data alterations suggested underreporting.

In conclusion, this paper highlights the visual analytics into survey design to thoroughly understand and effectively tackle the challenges associated with web-based data collection. By capitalizing on insights provided by paradata and leveraging advanced visualization tools, survey researchers can significantly improve both the accuracy and efficiency of self-administered surveys. This approach ultimately leads to more detailed, robust, and actionable data.

Keywords: paradata, survey quality, new tools, response behaviour, new data collection tools

1 Introduction

Achieving high-quality data collection through surveys can often be challenging, particularly during the development or revision phases. Errors can arise at various stages of the process, and strategies such as sample selection, response weighting, or a targeted focus on underrepresented groups during data collection are typically employed to mitigate this uncertainty.

When creating or modifying surveys, such as transitioning from telephone-based (CATI) to web-based formats (CAWI), user testing frequently plays a critical role. The aim of these tests is to pinpoint weaknesses in the questionnaires that might elicit poor responses from participants, and to prevent potential dropouts, an issue especially relevant in self-administered web-survey.

Various methods are currently employed to identify these weaknesses, ranging from quantitative experiments to more qualitative approaches like cognitive testing or focus groups. These methods are commonly resource-intensive, and the latter two can be particularly taxing on participants. The challenges mentioned can be encapsulated within the Total Survey Error (TSE) model, which addresses potential pitfalls from sample selection to the actual data collection execution, this article will specifically address "Nonresponse Error" aspect of the TSE model (Groves, et al., 2010).

The advent of self-administered web surveys (CAWI) has unlocked new data sources that can enhance the quality of surveys in line with the TSE model. These surveys are often favoured due to budgetary constraints and minimize respondent burden by allowing individuals to choose when to respond. However, this autonomy increases the demands on question formulation and design, since there are no interviewers present to aid respondents. The thematic content of the questions or their sequence might also lead respondents to discontinue the survey; generally, dropout rates are higher in self-administered surveys compared to those conducted via telephone. On the other hand, CAWI surveys generate electronic traces, known as paradata (Kreuter, 2013). This includes tracking keystrokes, mouse movements, and timestamping activities, which provide insights into the respondents' behaviour throughout the questionnaire.

This article demonstrates how we have utilized paradata to identify challenging questions, bottlenecks, underreporting, and the general flow of the questionnaires. We emphasize tools that visualize the data flow between questions or sections within the form, and underreporting.

2 Case and research question

In this article, we analyse the cross-border shopping survey, conducted by Statistics Norway (SSB), which has recently been converted from CATI to CAWI format. This survey was chosen because it is relatively simple compared to other surveys conducted by SSB. Respondents are asked about day trips abroad and the expenses related to the travel itself, as well as expenditures on goods or services purchased abroad.

The questionnaire is designed such that the number of questions about expenses increases based on the number of day trips abroad reported by the respondent. This means the questionnaire becomes significantly longer and contains repetitive questions for those respondents who have made multiple day trips. This has introduced uncertainty and questions about how this will affect responses from respondents who frequently travel on day trips abroad, despite the lack of a robust data foundation regarding respondent behaviour within the survey. We were aware that it might not necessarily be the question about the number of trips that is problematic, but rather how questions about expenses abroad might cause a respondent to either drop out or report a fewer number of trips. This suggests that the issue lies in the interaction among several questions or the user experience itself.

This led us to develop a visualization method that highlights bottlenecks and regressions in the questionnaire, or in other words, the user journey. This tool will help narrow down further analysis and provides indications of which parts of the questionnaire require further development.

To keep the analysis straightforward, we have chosen to group questions with the same theme into the following segments [1].

[1] *The order coincides with the sequence of segments in the survey.*

1. Intro
2. Trip
3. Number_trips
4. Accommodation
5. Purchases
6. Exit

3 Tools and data basis

During our work with visualization, we became acquainted with the Sankey diagram, developed by Riall Sankey. The diagram was originally created to illustrate flows from a specific source to a designated target, aimed at identifying energy losses in complex industrial systems (Schmidt, 2008). This underlying principle is also applicable in the analysis of CAWI-based questionnaires. Typically, respondents start at the beginning of the questionnaire and are guided through various segments based on their answers until the questionnaire is completed.

The Sankey diagram is ideal for visualizing how a large number of respondents navigate through the survey, effectively illustrating quantities, dropouts, and the direction of the flows. This enables the identification of bottlenecks and dominant flow patterns in the questionnaire, as well as unexpected transitions that may lead to further analyses or improvements of the questionnaire and thus the user experience.

In the analysis and visualization discussed in the article, we use paradata from the Cross-border Trade Barometer 2023, which includes 19,976 respondents and 503,210 registrations. The data originates from the Blaise system developed by Statistics Netherlands (CBS) and includes every keystroke, timestamped with information on where the click occurred. The data material is often overwhelming and contains much noise, as we do not know the context of the respondents' actions, such as prolonged time spent on one question. It is unclear whether the respondent spent all this time responding to the survey or if other factors caused them to leave the form for an extended period. Decisions on whether to include or exclude such respondents require domain knowledge and more thorough analysis. Additionally, the data requires extensive cleaning, as it also records much metadata, such as error messages that do not necessarily prevent respondents from continuing or completing the survey.

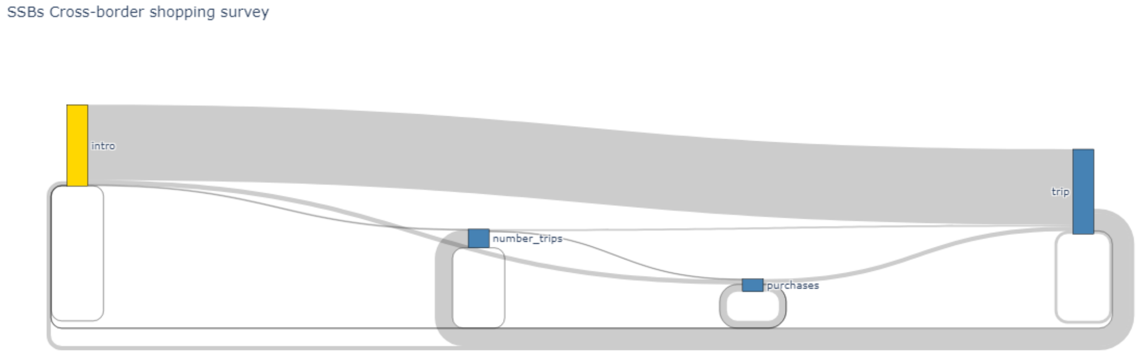
We have chosen to perform data wangling and cleaning using Python along with the various libraries that are available [2].

4 Data flow and analysis

As previously mentioned, we have specifically focused on potential underreporting due to the multiplying structure of the web form. It is well-known that such forms can reduce respondents' willingness to complete the survey and, in the worst case, result in measurement errors due to the length of the form (Peytchev, et al., 2017).

In the visualization below, we have limited the presentation to the journey from the start of the form to the registration of purchases abroad. This is done to keep the visualization clear and to highlight the flow.

Visualization 1: Flow chart



In the visualization, we observe clear transitions between different sections. Noteworthy is the relationship between "number_trips", which includes questions about the number of recorded trips, "purchases", which contains questions about reported amounts within various categories such as groceries or alcohol, and "trip", which deals with questions about whether the recorded trips were day trips abroad with or without an overnight stay. The first thing noted is the thickness of the connections between the sections; after the "trip" section, the thickness decreases, indicating that many respondents end the form if the trip abroad included an overnight stay, or that many respondents drop out. The remaining respondents move on to "*number_trips*", and the

[2] *The code is public through GitHub and can be used for own purpose and further development.*

line becomes even thinner before reaching "*purchases*", which contains questions regarding expenses related to the day trip.

We consistently observe that many respondents drop out. Additionally, loops around the various blocks indicate many recordings between questions within each section, suggesting a bottleneck in the form. This makes sense as respondents are asked to report expenses from the previous month, broken down into different categories of goods, which can be challenging to answer due to memory or lack of receipts. Most relevant for our analysis are the cross-links from "*number_trips*" and "*purchases*", too "*trip*". We see that there are links between these, meaning that respondents jump back to the question about having day trips abroad from the questions about the number of trips and expenses. In summary, we observe dropouts during the completion, bottlenecks, and back-jumping to questions about trips abroad. By using this method, we have identified weaknesses in the form, and we have limited the analysis to focus on parts of the form concerning day trips abroad.

To investigate whether underreporting occurs in the question about day trips, we have examined whether the same question is recorded with different timestamps for each respondent. We then compare the answers between the first and last registration. If the deviation is other than zero, it indicates that the respondent has changed their answer afterwards. In the table below, we see that a third of the respondents who had more than one registration on the question about day trips abroad had differing answers between the first and last registration.

Table 1: Number of answer changes

	Number of changes	**Percentage**
Day trips Abroad	482	33,7%
Other questions in the survey	950	66,3%
Sum	**1 432**	**100,0%**

Changes recorded in the survey may be misleading because every keystroke is logged, and some keystrokes may be accidental. Therefore, it is prudent to focus on the last recorded value for each question. In the question about day trips, a "yes" is recorded as the value 1, while a "no" is recorded as 2.

This means that a difference of 1 indicates that the first recorded value was 2 and the last value was 1, suggesting that the respondent changed their answer from no day trips to one or more day trips abroad. Conversely, a difference of -1 indicates that the initial registration was one or more day trips, but changed to no day trips abroad, which could suggest underreporting.

By analysing only, the values associated with the question about day trips, the data show that 61% of all recorded changes were to a lower value, indicating a certain degree of underreporting.

Tabel 2: number of value changes

	Number of observations	**Prosent**
Reduced value (-1)	294	61,0%
Increased value (1)	188	39,0%
Sum	**482**	**100,0%**

The use of Sankey diagrams for visualizing data flow has shown us how we can observe respondents' movements through the questionnaire. We have uncovered several aspects of the questionnaire that were partly known before and require further attention. Specifically, we have identified bottlenecks at questions about reported amounts, and we have observed underreporting related to the reporting of day trips abroad.

The most critical finding is that it is not necessarily the individual questions that lead respondents to change their answers, but rather the interplay between questions about the number of trips and the reporting of expenses that influences the answers to questions about day trips. This insightful discovery underscores the importance of considering how different parts of a questionnaire affect each other and the respondents' perception and completion of the survey.

5 Conclusion

The transition to CAWI has highlighted significant challenges in survey methodology, particularly regarding respondent behaviour and data integrity. The findings from the use of paradata in the Cross-border shopping survey reveal critical insights into the dynamics of questionnaire completion, notably the impacts of question sequence and formulation on respondent dropout and data reporting. The use of visual tools like Sankey diagrams has proven invaluable in pinpointing bottlenecks and understanding the respondent journey, leading to more informed decisions about survey design.

The research underscores the importance of continuous evaluation and adaptation of survey methodologies to better capture accurate and comprehensive data. Future efforts should focus on refining these visual tools and incorporating more dynamic methods of engagement to reduce dropout rates and improve the quality of data collected. This will be crucial for harnessing the full potential of CAWI and ensuring that surveys are both cost-effective and robust in their findings.

6 Bibliografi

A. Regula, H., & Jerald G., B. (1981). Effects of Questionnaire Length on Response Quality. *The Public Opinion Quarterly, vol. 45, no. 4*, pp. 549–559.

Groves, R. M., & Lyberg, L. (2010). TOTAL SURVEY ERROR: PAST, PRESENT, AND FUTURE. *Public Opinion Quarterly, 74(5), 849–879*. doi: https://doi.org/10.1093/poq/nfq065

Kreuter, F. (2013). *Improving surveys with paradata : analytic uses of process information.* John Wiley & Sons, Inc.

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. *Section on Survey Research Methods – JSM 2010*, pp. 282-296.

Peytchev, A., & Peytcheva, E. (2017). Reduction of Measurement Error due to Survey Length: Evaluation of the Split Questionnaire Design Approach. *Survey Research Methods*(11(4)).

Schmidt, M. (2008, Mars 19). The Sankey Diagram in Energy and Material Flow Management. *Journal of Industrial Ecology, 12*, pp. 82-94.

Schmidt, M. (2008). The Sankey Diagram in Energy and Material Flow Management. *Journal of Industrial Ecology*(12).