Ref. Ares(2024)4663643 - 28/06/2024




Smart Survey Implementation

Grant Agreement Number: 101119594 (2023-NL-SSI)

Work package 2: Research Methodology

Deliverable M14: Smart baseline stage

Version 2.2 (submitted version)

Prepared by:
Hannah Bucher & Florian Keusch University of Mannheim, Germany
Johannes Volk, Lasse Häufglöckner, Karen Blanke, DESTATIS, Germany
Claudia de Vitiis, Fabrizio de Fausti, Francesca Inglese, Marco Terribili & Monica Perez, ISTAT, Italy
Theun Pieter van Tienoven, Vrije Universiteit Brussel, Belgium
Patrick Lusyne, Statistics Belgium, Belgium
Danielle McCool, Peter Lugtig & Bella Struminskaya, Utrecht University, the Netherlands
Anne Elevelt, Jelmer de Groot, Maaike Kompier, Barry Schouten, Jonas Klingwort & Janelle van den Heuvel, Statistics Netherlands, the Netherlands
Julie Solard, Simon Quentin, INSEE, France
Nina Berg & Aina Holmøy, SSB, Norway

ESSnet co-ordinator: Remco Paulussen
Workpackage Leader:
Peter Lugtig
Utrecht University
p.lugtig@uu.nl

Disclaimer: Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Eurostat. Neither the European Union nor the granting authority can be held responsible for them.

Index

Introduction

The goal of the WP2 'Methodology' workpackage of the Smart Survey Implementation project (SSI) is to find out what general methodological elements trusted smart surveys should have so that they can be used in statistical production by European NSIs. Each task focuses on either an 'opportunity' or 'threat' that was identified in the TSS I framework and pilot recommendations for smart surveys. The four subtasks are:

1. The successful **recruitment** of participants for smart surveys.
2. Using **machine learning** to improve Human-Computer Interaction in smart surveys.
3. Respondent involvement and **human-computer interaction** in smart surveys.
4. Integrating smart surveys with traditional survey methods by estimating the mode **effect.**

We refer to deliverable M6 (Review stage) for a discussion of learnings from past findings from projects conducted in the context of the European Statistical System and the wider academic context with regards to these four key challenges. In this deliverable, findings from the M6 deliverable are in places summarized when this is necessary to understand the current deliverable, in which we explain how we want to answer the challenges identified in the M6 deliverable within the Smart Survey Implementation project.

Of central importance are several large and small field tests. A series of large field tests conducted in Q1 – Q4 of 2024 will focus on understanding the **recruitment and the mode effect**, and a series of small tests conducted throughout 2024 and early 2025 will focus on **human-computer interaction** and **machine learning**.

The small tests at the core of testing UI/UX in task 2.3 have a focus on the implementation of smart surveys in the context of understanding household expenditure in the Household Budget Survey (HBS) and the Time Use Survey (TUS). The goal of these smaller experiments is to technically test the Machine Learning standards developed in task 2.2, to test the HCI features of smart surveys in task 2.3, and to technically test some of the microservices developed in Workpackage 3 that are completed by the time these tests are scheduled.

The small tests are qualitative, and they will use as many respondents as necessary before saturation is reached (e.g., n=~20). Samples here are taken from existing surveys (follow-ups of respondents from other surveys), online access panels or other volunteers. The goal here is not to draw inferences to the general population but to include a diverse set of respondents (e.g., in terms of age, internet experience), so that the Human Computer Interaction is tested for different types of potential respondents. Tests are foreseen in all countries in the consortium and carried out throughout the project, scheduled in close alignment with WP 3 (microservices). Chapter 3 of this deliverable describes the small tests in some detail, as well as the test protocols used to test the respondent involvement and human-computer interaction of smart surveys

The large tests aim to answer the question of how respondents can be successfully be recruited into smart surveys (task 2.1) and how to integrate smart surveys with traditional surveys (task 2.4). The large field tests will be conducted in Norway (HBS), France (TUS & HBS), Belgium (TUS), Germany (HBS), Italy (TUS) and the Netherlands (HBS). Norway and France will use a smart survey app which was self-developed, Germany, Belgium and Italy will use MOTUS, and the Netherlands will use the HBS platform. All countries will use the general population as the target population and draw fresh samples to conduct the field test following a general design, where some key elements of the field tests are shared across the countries. Respondents are recruited using an offline method (e.g.

recruitment via interviewers or postal mail) and are based on large probability samples. This allows data from multiple countries to be pooled in the analysis, increasing statistical power and allowing for analyses into recruitment effects for smaller subgroups in the general population. At the same time, it also allows for the comparison of country-level differences in, for example, the success of particular recruitment strategies.

The exact design of the experiments around communication materials and **recruitment** (task 2.1) is worked out in Chapter 1 of this deliverable. Chapter 4 focuses on the design and experiments carried out to inform the **mode effect** of smart surveys (task 2.4).

Table 0.1: Overview of Large Field Tests in Individual Countries

Country	**Survey**	**Field period**	**Sample size (gross)**	**Task(s)**
Germany	HBS	September 2024 (2 weeks)	7.000	2.1
Belgium	TUS	March/April 2024	6.000	2.1/2.4
Netherlands	HBS	September/October 2024	1.600	2.1/2.4
Norway	HBS	April/May 2024 (6 weeks)	2.000	2.1
France	TUS & HBS	April 2023/ May 2024 (2 months each)	2100+2400	2.4
Italy	TUS	End October/November	5.000	2.4

Notes: HBS: Household Budget Survey. TUS: Time Use Survey

Chapter 2 follows closely on the M6 deliverable and focuses on the role of Machine Learning in processing smart data that is used in the field tests for the Time Use Survey (geolocation data) and Household Budget Survey (pictures of receipt). Chapter 2 bridges the more technical work of WP 3, in which microservices are developed, to the work of task 2.3 (human computer interaction). This deliverable presents partly work in-progress on how sensor data are being processed in the microservice using machine learning and presented back to the respondent. As the microservices are at this stage (June 2024) still under development, a final test of how good the machine learning models work and how processed data can be fed back to the respondent will be part of the M24 deliverable, in which information from the small and large field tests can be used to evaluate the quality of the machine learning models used in smart surveys for TUS and HBS, and the end-to-end process as a whole.

Utrecht, 27 June 2024

The authors.

1. Enhancing Recruitment Strategies for Smart Surveys in Official Statistics: Experiments and Insights from Work Package 2.1

1.1 Background

The emergence of smart surveys that combine the use of asking questions (surveys through self-report) with smart features collected via sensors on smartphones, wearables and other devices presents an opportunity to enhance the efficiency and effectiveness of official statistics across Europe. However, the success of smart surveys highly depends on participant engagement. The objective of Work Package 2.1 (WP2.1) within the SSI project is to optimize recruitment strategies for these surveys, aiming to significantly increase participation rates and ensuring a sample composition that allows to draw inferences to the entire population based on the realized sample for official statistics.

1.2 Challenges in Smart Survey Participation

As detailed in Deliverable M6, prior research has identified two main categories of factors influencing the willingness to participate in smart surveys. The first category includes study characteristics under the control of researchers, such as the design of the participation invitation, the involvement of interviewers, the presence of a landing page, and the use of incentives. The second category comprises respondent characteristics, like technical knowledge and privacy concerns, which are beyond direct researcher influence but significantly affect participation (Keusch et al., 2023). However, these respondent characteristics can be highlighted and addressed within the controllable elements of survey design, effectively mitigating these challenges by making them salient to participants (Groves et al., 2000). For instance, individuals with lower digital literacy or heightened privacy concerns are often hesitant to participate in smart surveys. Addressing these hesitations is one of the major challenges in recruiting respondents for smart surveys, as it requires a nuanced understanding of how survey design features can facilitate smart survey participation. By systematically addressing these aspects within the experiments conducted in WP2.1, we aim to refine the recruitment process for smart surveys. The ultimate goal is to ensure higher response rates and participation from a sample that mimics the population, thereby producing reliable and valid data for official statistics in Europe.

1.3 Research Objectives

In deliverable M6, we conducted an extensive review of existing literature to pinpoint effective survey design features that could mitigate barriers to participation in smart surveys. This

comprehensive analysis led to the formulation of three key objectives that will be empirically tested in our upcoming experiments:

1. **Enhancing Trust and Alleviating Privacy Concerns:** Our first objective focuses on increasing trust in the data collection organization and alleviating privacy-relatedconcerns during the invitation process. This involves crafting invitations that transparently communicate data handling practices and privacy safeguards to reassure respondents.
2. **Tailoring Interfaces for Diverse Digital Literacy Levels:** The second objective aims to adapt survey interfaces and processes to accommodate respondents with varying degrees of digital literacy. This effort is intended to minimize perceived complexity and effort, making the survey more accessible and less intimidating for all potential participants.
3. **Applying Proven Engagement Strategies:** The third objective involves adapting successful strategies from traditional survey methodologies—such as the Tailored Design Method (Dillman et al., 2014) and principles from Social Exchange Theory (Blau, 1964; Dillman et al., 2014) —to smart surveys. These strategies are intended to enhance engagement by making the survey experience more personalized and reciprocally beneficial.

We plan to implement these features at different stages of the recruitment process, including the survey invitation, the app downloading platform, and follow-ups with individuals who did not initially respond. The experiments within WP2.1 will cover all these stages, thus providing a comprehensive approach to recruiting respondents for smart surveys. Figure 1 gives an overview of the approach taken in the experiments of WP2.1.

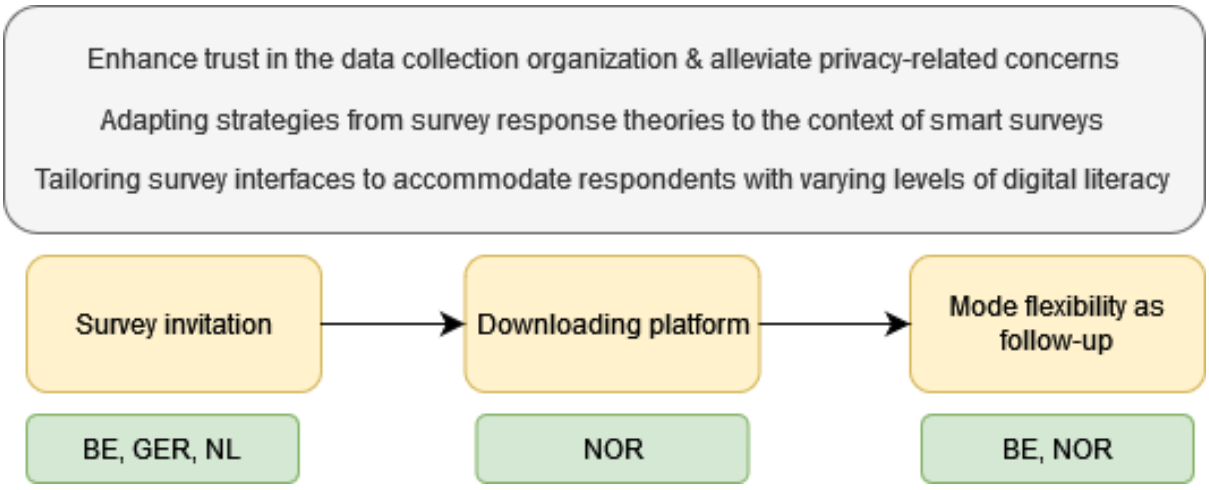


Figure 1.1: Scope of the experiments conducted within WP2.1

1.4 Overview of the experiments

The experiments conducted across Belgium, Germany, the Netherlands, and Norway collectively aim to refine and enhance the recruitment strategies for smart surveys. In Belgium, the focus is on assessing how invitation design and the option of a PAPI follow-up can influence response rates and sample composition. Germany's experiments explore theimplications of disclosing smart survey features and data collection efforts on respondent engagement. Meanwhile, the Netherlands investigates the effectiveness of using interviewers and prenotification letters to boost participation, especially among hard-to-reach groups. Norway's approach examines the influence of trust and familiarity with the download site on the willingness to participate, testing the impact of downloading apps from government sites versus standard app stores. Furthermore, the Norwegian field experiment includes an embedded study to investigate how offering a CATI (Computer-Assisted Telephone Interview) follow-up option influences response rates and sample composition.

Together, these experiments are designed to identify effective approaches that address digital literacy, privacy concerns, and the general willingness to participate in surveys. By integrating traditional survey methods with innovative digital approaches, these studies seek to optimize response rates and ensure diverse respondent inclusion in the era of smart surveys. This overarching goal supports the broader aim of adapting survey methodologies to contemporary challenges and opportunities in data collection. Table 1.1 gives an overview of the experimental design of each country. Materials used in the fieldwork of the countries can be found in Appendix B-G.

Table 1.1: Overview of the experiments conducted in the large fieldtests and focus in participants recruitment for smart surveys.

Country	Research question	Theoretical background	Hypotheses	Design
BE	**R1:** How does the overall design of the invitation letter affect the response rate?	**Tailored Design Method** (Dillman et al., 2014): The use of clear, concise messaging and the incorporation of user-friendly design elements in the survey invitation should increase response rates.	**H1:** Incorporating user-friendly elements in the design of the invitation will significantly improve response rates in a smart survey.	*Invitation Letter* Traditional / User-friendly *PAPI follow-up*: Yes / No
	R2: How does the inclusion of a follow-up PAPI mode affect the response rate in a smart survey?	**Digital Literacy**: Previous research indicates that respondents with lower digital literacy are less likely to participate in smart surveys (Jäckle et al., 2019; Keusch et al., 2022; Revilla et al., 2019; Struminskaya et al., 2020, 2021; Wenz & Keusch, 2023). By offering a PAPI mode, these individuals might be more inclined to participate. **Privacy Concerns**: Individuals with high privacy concerns are often hesitant to engage in surveys that involve smart data collection methods (Keusch et al., 2020; Oyibo & Pelegrini Morita, 2022; Revilla et al., 2019; Wenz et al., 2019; Wenz & Keusch, 2023). Offering a PAPI mode can help mitigate these concerns by potentially raising participation rates by assuring respondents of a more secure and traditional data collection approach.	**H2.1:** Offering a PAPI mode positively influences the overall response rate in surveys. **H2.2:** The inclusion of a PAPI mode improves the sample composition of surveys by attracting also respondents with lower digital literacy and higher privacy concerns.	
GER	**R1:** How does highlighting the smart features of a survey in the invitation influence its response rate?	**Simplifying the data collection process:** Utilizing a smartphone's camera simplifies the data collection process, making it more convenient for respondents. This method reduces the effort required to participate, thereby potentially lowering the overall burden of survey participation. **Privacy Concerns:**	**H1.1**: Stressing the smart features of the survey in the invitation positively affects the response rate. **H1.2:** Stressing the smart features of the survey in the invitation will	*Microservice mention in invitation* Yes / No *Effort of data collection mention*: No

		Mentioning the smartness of the survey in the invitation may negatively impact the overall response rate due to heightened privacy concerns among respondents (Keusch et al. 2020; Oyibo and Pelegrini Morita 2022; Wenz and Keusch 2023; Revilla, Couper, and Ochoa 2019; Wenz, Jäckle, and Couper 2019b). Specifically, stating that the app will access the smartphone's camera could increase awareness and apprehension regarding data protection issues.	negatively impact the response rate	Yes [redacted]
	RQ2: To what extent does highlighting the benefits of the scanning function lead to increased usage of this feature?	**Social exchange theory:** If the benefits of participation outperform the costs, respondents decide to take part in the survey (**Blau, 1964**). Related to the usage of smart features of surveys, this could also apply to the usage of allowing the survey access to the camera, since it dramatically reduces the costs of data collection for the respondents.	**H2** Emphasizing both the effort required to complete the survey and the benefits of the scanning feature will lead to higher usage of the scanning function by individuals, compared to merely mentioning the scanning function without stressing the effort involved.	
NL	**RQ1:** Does the usage of Interviewers in the course of the recruitment positively affect response rates?	**Providing assistance in downloading and installing the app:** Interviewers could help individuals with lower digital literacy (Jäckle et al. 2019; Keusch, Wenz, and Conrad 2022; Revilla, Couper, and Ochoa 2019; Struminskaya et al. 2020; Struminskaya et al. 2021; Wenz and Keusch 2023). **Addressing privacy concerns:** Giving respondents the opportunity to discuss potential privacy concerns with the interviewers, may help overcoming burden related to privacy concerns as it increases the trust in the survey agency. **Willingness to participate:** The personal contact inherent in interviews offers several advantages that can improve the quality and completeness of the data collected, such	**H1.1:** The use of interviewers in the recruitment process results in higher response rates compared to conditions where interviewers are not used. **H1.2: The use of interviewers in the recruitment process results in a sample that better mimics the population compared to**	**RECRUITMENT PROCESS** Letter | Prenotification & letter | Letter & interviewer **To investigate the effectiveness of prenotifications and interviewers in the recruitment process, the design in the Netherlands uses three different subsamples: first, a subsample that mirrors the target population; second, a subsample consisting of hard-to-reach respondents stratified by age and country of birth; and third, a subsample composed of individuals with high participation probabilities.**

		as motivate and encourage individuals to participate in the survey (Groves et al., 2009; Heerwegh & Loosveldt, 2008).	**conditions where interviewers are not used.**	
	RQ2: Does the usage of announcement letters in the course of the recruitment positively affect response rates?	**Social exchange Theory:** Prenotification serves as a preliminary engagement that reduces potential respondents' perceived costs and enhances perceived rewards **(Dillman et al., 1976; Leeuw et al., 2007; Tourangeau, 2017)**.	**H2**: The use of announcement letters in the recruitment process results in higher response rates compared to conditions where announcement letters are not used.	
	RQ3: Does the usage of interviewers and announcement letters help to reach respondents with low participation probabilities?	Social Exchange Theory (Blau, 1964): Interviewers and prenotifications can build trust and perceived value of participation, overcoming reluctance. Tailored Design Method (Dillman et al., 2014): adapting communication strategies to the specific circumstances and preferences of hard-to-reach groups ensures that outreach efforts are perceived as relevant and respectful, thereby increasing engagement.	**H3:** The usage of announcement letters and interviewers is most successful for the recruitment of hard-to-reach respondents.	
NOR	**RQ1:** Does the type of download site for the app influence RR?	**Trust/privacy concerns:** Trust in organization has a positive impact on willingness to participate **(Wenz & Keusch, 2023)**.	**H1:** Higher response rates are associated with downloading the app from a governmental site (AltInn).	*Downloading platform* AltInn / App store *CATI follow-up* Yes No

1.5 Analysis Strategy Overview

The analysis strategy for evaluating the experiments from smart surveys conducted across various countries focuses on two primary aspects: response/contact rates and sample composition. The objective in assessing response and contact rates is to compare the effectiveness of different experimental conditions within and across countries. We will calculate RR1 and RR2 adhering to the American Association for Public Opinion Research (AAPOR) standards (AAPOR, 2016). This allows for a detailed examination of which strategies are most successful in enhancing participation rates. Additional statistical measures, such as R indicators are employed to provide a deeper understanding of response distributions and disparities within the collected data. With this approach, we aim to estimate the likelihood of response among different subgroups within the target population.

For sample composition, the analysis will look at the performance of the survey samples in depicting the target population in each country. The approach includes a first analysis using all available demographic and socio-economic variables specific to each country, providing insights into the country specific performance of the realized survey samples. For a European perspective, a between-country comparison will be conducted using only the variables available across all surveys, ensuring fair and robust comparisons. We will further use the Gini coefficient, typically used to measure income inequality, adapted for assessing the balance of the survey sample compared to the target population.

The methodology involves a two-step analytical process: starting with an in-depth analysis of each country's data using all available variables to understand country-specific dynamics and moving to a broader, international comparison using shared variables to evaluate the generalizability and effectiveness of survey strategies. This comprehensive approach allows for tailored strategies as well as the development of broader insights into the effectiveness of survey methodologies in the context of European official statistics. Figure 2 gives an overview of the analytical approach employed.

1.6 Conclusion

Workpackage 2.1 of the SSI project encompasses a comprehensive effort to advance recruitment strategies for smart surveys across Europe. By integrating innovative digital methods with traditional survey techniques, the experiments conducted in Belgium, Germany, the Netherlands, and Norway aim to tackle significant challenges associated with digital literacy, privacy concerns, and the overall willingness to participate in smart surveys.

Throughout these experiments to enhance participation rates, a variety of approaches are to be explored, from modifying invitation designs to introducing follow-up modes such as PAPI and CATI.

The learnings from Work Package 2.1 not only contribute to the field of survey methodology but also set a precedent for future statistical data collection efforts in Europe, aiming to harness the full potential of smart technologies while addressing the concerns and preferences of the public.

References

AAPOR. (2016). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. The American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf

Blau, P. M. (1964). Exchange and power in social life. Wiley.

Dillman, D. A., Gallegos, J. G., & Frey, J. H. (1976). Reducing Refusal Rates for Telephone Interviews. Public Opinion Quarterly, 40(1), 66–78. https://doi.org/10.1086/268268

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method (4th edition). Wiley.

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (Hrsg.). (2009). Survey methodology (2. ed). Wiley.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-Saliency Theory of Survey Participation. Public Opinion Quarterly, 64(3), 299–308. https://doi.org/10.1086/317990

Heerwegh, D., & Loosveldt, G. (2008). Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality. Public Opinion Quarterly, 72(5), 836–846. https://doi.org/10.1093/poq/nfn045

Jäckle, A., Burton, J., Couper, M. P., & Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: Coverage and participation rates and biases. Survey Research Methods, Vol 13, 23-44 Pages. https://doi.org/10.18148/SRM/2019.V1I1.7297

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2023). Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey. Sociological Methods & Research, 52(2), 841–878. https://doi.org/10.1177/0049124120914924

Keusch, F., Struminskaya, B., Kreuter, F., & Weichbold, M. (2020). Combining Active and Passive Mobile Data Collection: A Survey of Concerns. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Hrsg.), Big Data Meets Survey Science (1. Aufl., S. 657–682). Wiley. https://doi.org/10.1002/9781118976357.ch22

Keusch, F., Wenz, A., & Conrad, F. (2022). Do you have your smartphone with you? Behavioral barriers for measuring everyday activities with smartphone sensors. Computers in Human Behavior, 127, 107054. https://doi.org/10.1016/j.chb.2021.107054

Leeuw, E. D., Callegaro, M., Hox, J., Korendijk, E., & Lensvelt-Mulders, G. (2007). The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis. Public Opinion Quarterly, 71(3), 413–443. https://doi.org/10.1093/poq/nfm014

Oyibo, K., & Pelegrini Morita, P. (2022). Factors Influencing the Willingness to Download COVID-19 Contact Tracing Apps: The Moderating Effect of Persuasive Design and Smartphone Usage Experience. Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, 11(1), 163–169. https://doi.org/10.1177/2327857922111033

Revilla, M., Couper, M. P., & Ochoa, C. (2019). Willingness of Online Panelists to Perform Additional Tasks. Methods, data, 29 Pages. https://doi.org/10.12758/MDA.2018.01

Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges. Social Science Computer Review, 089443932097995. https://doi.org/10.1177/0894439320979951

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2021). Understanding Willingness to Share Smartphone-Sensor Data. Public Opinion Quarterly, 84(3), 725–759. https://doi.org/10.1093/poq/nfaa044

Tourangeau, R. (2017). Mixing Modes: Tradeoffs Among Coverage, Nonresponse, and Measurement Error. In P. P. Biemer, E. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Hrsg.), Total Survey Error in Practice (1. Aufl., S. 115–132). Wiley. https://doi.org/10.1002/9781119041702.ch6

Wenz, A., Jäckle, A., & Couper, M. P. (2019). Willingness to use mobile technologies for data collection in a probability household panel. Survey Research Methods, Vol 13, 1-22 Pages. https://doi.org/10.18148/SRM/2019.V1I1.7298

Wenz, A., & Keusch, F. (2023). Increasing the Acceptance of Smartphone-Based Data Collection. Public Opinion Quarterly, nfad019. https://doi.org/10.1093/poq/nfad019

2. Machine Learning in smart surveys: insights from workpackage 2.2

This section presents developments and considerations on ML-methods for processing sensor data in the contexts of receipt scanning and geotracking.

As already highlighted in deliverable 2.1, for receipt scanning data the experience gained by the NSIs and in European projects in the context of smart surveys for official statistics is at a more advanced stage. Also, in the timeline of the SSI project the development of the receipt scanning microservice came first and therefore the discussion on methodological issues is more mature. The development of the methodology for the Geotracking microservice is still ongoing at the moment of writing. Nonetheless, many reflections have already been made on both sides.

Sections 2.1 and 2.2 address respectively receipt scanning and geotracking.

2.1 ML for receipt scanning and COICOP classification

Semi-automated handling of shopping receipts is the smart case study considered for an app-assisted Household Budget Survey (HBS). Two forms are distinguished: paper receipts that are scanned by respondents and e-receipts that are uploaded by respondents. The two forms correspond to two different categories of smart features in the taxonomy (see deliverable 4.2 of WP4). Uploading e-receipts is seen as a form of data donation. It has not yet been considered in the Smart baseline stage of SSI. Here, the focus is on scanning printed receipts.

The handling of paper receipts consists of three main steps that are sketched in deliverable 3.1 of WP3. The first step is in-device and comprises of taking a picture including respondent-interaction on the quality of a scan/photo. It is a platform/solution-specific step. However, UI-UX aspects are tested in SSI (see the WP2.3 sections in deliverable 2.2 of WP2). The second step is text extraction and document understanding and implies the conversion of pixels to meaningful text units. The third step is classifying products derived in the second step to the formal COICOP classification employed within the ESS. Steps two and three are microservices that can be used by any platform/solution. These two steps have clear methodological/data science choices that are discussed here. Performance of the two services needs to be evaluated in time and space (i.e. countries) and needs to be translated back to UI-UX. This is the task of WP2.2 and WP2.3.

2.1.1 Receipt text extraction

Receipt text extraction has been a focal point of WP3 in stage Smart Baseline and a microservice has been developed. The details can be found in deliverable 3.2 of WP3.

2.1.1.1 The receipt text extraction service

The text extraction of printed receipts itself comprises of three substeps: pre-processing scans (orientation, edge detection, etc.), identification of text boxes and classification of text boxes. Underlying the classification of text boxes is machine learning. Text boxes can have a range of types (see again deliverable 3.2) such as total amount, date, shop name, product, price. The SSI microservice assumes that all identified text boxes are classified based on a pre-trained model. The model training is done prior to fieldwork by annotating a set of (training) receipts. Nonetheless, as a backup, it is up

to a platform/solution to also ask respondents for receipt descriptives such as total amount or shop name.

The output of the microservice consists of:

- A vector of text box locations
- Per text box:
    - an ID
    - the extracted text
    - a vector of classification predictions mapping to the categories chosen by WP3
    - In case type is product or price: the corresponding product ID or price ID
- OCR performance score A list of flags for the detection of each mandatory text type (date, shop name, ≥1 product, ≥1 price, total amount)

It must be noted that text box classification predictions do not necessarily reflect true probabilities. Texts or text types that have never been observed before may be erroneously get high predictions.

2.1.1.2 Country-specific decisions

What may be country-specific:

- Language (some countries in addition have multiple languages, e.g. CH and LU)
- Receipt lay-out
- Diversity and balance of receipt layouts
- Punctuation
- Currency
- Range of text box types
- OCR performance

Given that many machine learning features in the training of the text box classification may change, per country/language a new model is trained. The speed at which model training converges is yet unknown but will be evaluated using field tests in DE and NL. Convergence will depend on feature importance and the variety in feature values. Consequently, it will vary per country.

What countries need to do:

- Identify what features are different in practice
- Evaluate these differences against feature importances identified for BE-DE-NL
- Create a balanced set of annotated receipts. Recommendations on the minimal train set size have yet to be determined in the last SSI project stage

2.1.1.3 Comparability in time

What may change in time:

- Receipt lay-out
- Range of text box types
- Receipt texts

Dynamics are conjectured to be modest in size. The most imminent change is from printed to digital receipts. Digital receipts may themselves come in different formats.

The actual dynamics in receipt features are unknown but will be evaluated in the last SSI project stage Smart Advanced.

2.1.1.4 From output to input

The output of the receipt text extraction service may be input to the UI of a platform/solution. The UI can follow one of the following rules to initiate respondent interaction:

- Are all mandatory text box types present?
- Is there a conflict between manual user input and extracted text?
- Is a text box classification prediction below a specified threshold, (perhaps further elaborated for the mandatory types)?
- Does every product have a price and vice versa?
- Is the number of products above a certain threshold (pointing at a long receipt)?

2.1.2 Product classification

Product classification investigations started within SSI in project stage Smart Baseline. They led to a multi-step procedure that is introduced in deliverable 3.2 of WP3. In project stage Smart Baseline only a basic version is embedded in a microservice. In stage Smart Advanced it will be perfected based on fieldwork data from DE and NL.

2.1.2.1 Classification pipeline

Classification follows text extraction. The output of the receipt text extraction service, plus any additional information supplemented by respondents, are input to the classification service. The classification service has as input:

- Shop name
- Product text (alleged)
- Product price (alleged)
- Text box type prediction
- OCR performance score

Shop name and product text are features. Product price may be included as a feature. Text box prediction and OCR performance score are optional features.

The classification pipeline assumes a set of shops/shopping chains for which either product texts are available or a ML model has been trained. Let this set be . The pipeline is a four-step procedure:

1. If the detected shop is in , then proceed, otherwise go to 2
    a. Apply string matching to the known and labelled set of products for shop . If found with acceptable matching distance, then stop.
    b. Apply string matching to the known and labelled set of products for other available shops . If found with acceptable matching distance, then stop.
    c. Apply the pre-trained model for shop . If the largest classification prediction is acceptable, then stop.
2. Apply pre-trained models of other shops. If acceptable prediction, then stop.
3. Forward the product text to manual classification.

The service includes 1a, 1b, 1c and 2. In order to perform the corresponding steps, it assumes a list of product texts per shop and compiled, pre-trained models for all shops for the country in which the service is employed.

It has not yet been decided what string-matching distance will be used. The current @HBS app uses a Jaro-Winkler distance.

A wide range of ML methods is available. Whether a single or an ensemble method is used will be decided in the last project stage of SSI.

The output of the service is:

- step n which prediction is based on (matching, ML of shop, ML of other shop, no matching)
- a vector of classification predictions at COICOP 5-digit, if applicable
- matching distance, if applicable

It has yet to be decided whether the service will also provide classification predictions for higher COICOP levels. If so, pre-trained models need to be hierarchical.

2.1.2.2 Country-specific decisions

What is country-specific:

- Set of shops for which product lists can be made available
- Set of shops for which a pre-trained ML model can be trained
- Type of training data

Training data may come from one of three sources:

1. Manually annotated receipts
2. Printed receipt texts linked to EAN/GTIN provided by shops
3. Scanner transaction product descriptions linked EAN/GTIN provided by shops

Sources 2 and 3 come with COICOP links prepared independently of HBS, usually by CPI departments.

The availability of type and size of training data will strongly impact performance of the pipeline. Again, it must be noted that trained ML models may give spuriously high predictions. This happens when product texts fall outside the body of training data. This may happen, for example, when a new type of item is added to a shop(e.g. clothing sold in a foodstore or a garden center) or when shops have a completely different lexicon (e.g. relatively short strings with abbreviations or brand names).

To date, investigations and training have been limited to supermarkets. Here, it must be noted that supermarkets may have types of alien products (clothing, electronics, etc). Classification accuracy for such products tends to be low due to their relatively small number and large diversity.

Countries need to decide:

- What types of shops handle semi-automatically
- What data sources to use for training
- If and how to set up direct contact with shops on their receipt text grammar
- Given a set of train data, how to monitor convergence
- What to do with branch-alien product types
- What thresholds to set in the pipeline
- What to do with products that cannot get classified (involve participant or NSI manual post-processing).

SSI will provide guidelines in stage Smart Advanced on how to make these decisions.

2.1.2.3 Comparability in time

Explorations and research in stage Smart Baseline clearly show that dynamics in receipt texts is large and influential for supermarkets. Earlier studies point at decreasing accuracy of around 1 to 1.5% per month.

What can change in time:

- The EAN/GTIN code of a product changes, implying that a link to COICOP needs to be re-established. Often the consequence of a change in metrics or look.

- The EAN/GTIN code is the same, but the product text changes
- New products are introduced
- Products are taken out
- Products may have seasonal fluctuations

A country/NSI has to decide:

- At what frequency to retrain
- If and how to set up a communication strategy with the main shops/chains on their text revision strategy

SSI will provide recommendations on the adaptive learning strategy.

2.1.2.4 From output to input

It has yet to be decided whether product classification will be communicated with respondents, and, subsequently, whether it may lead to respondent interaction. In addition, and as an alternative, the decision to go manual may be made. Again, certain rules may be used:

- Did any of the steps find an acceptable category?
- Is a COICOP classification prediction or matching distance is below a certain threshold?
- Was the prediction based on the ML model of another shop?

2.2 AI/ML for Geotracking data

2.2.1. Geotracking data in Smart Surveys

Employing location tracking data is potentially very useful for respondents as a framework to construct diaries concerning both travel and daily activities. Processing location tracking data to predict the behaviour of respondents requires advanced AI-ML exploiting different smart features and external auxiliary information.

The predicted variables can be used as response variables or as tentative data to be presented to the respondent. In both cases the predictions should be managed in the UI.

This is not always easy as there is no proof of the correctness of predictions based on GPS tracks at the beginning of the process. Prompted recall surveys involve asking respondents to check and, if necessary, correct their predicted variables, sometimes combined with an invitation to provide additional information that cannot be captured by GPS or other technologies. The combined use of GPS and a prompted recall requires less effort because respondents only have to check and correct the predicted data, rather than providing complete diaries of (out-of-home) activities and trips. This solution may still not be satisfactory from a burden reduction perspective. Therefore, alternative ways to improve prediction are being sought in the project.

As pointed out in Deliverabe 2.1, while the use of geotracking data has been explored by NSIs for the prediction of travel variables (travel mode and travel purpose, starting with the segmentation of GPS data in stops and trips), the methodological research for the prediction of the daily activities performed at the located places, as classified for the HETUS purpose, is more oriented to commercial purpose than to official statistics needs.

What is in common to Travel surveys and Time use surveys is the first processing step, the segmentation in stop and track, which provides the input variables for the algorithms devoted to predict the survey variables. A fundamental part in this process is the addition of auxiliary information

provided by a third party, such as map services, which provide information about the context around the GPS points and the Point of Interest (POIs) located nearby the GPS points.

The following sections will describe the current state of developments and discuss the main methodological issue.

2.2.2. Approach to predict survey variables based on GPS data

- GPS tracking data provides information on longitude, latitude, date, time, speed, altitude and direction of movement. This data can be used for mobility research to determine mode and purpose of travel, but also for other research domain (TUS survey), to predict daily activities performed at the located places.

Travel mode detection based on GPS tracking data is challenging, especially when the data is unlabelled, i.e. there is no information about the mode used during a trip. To detect transport modes, first trips and stop have to be extracted from the traces of GPS location. Spatial and temporal information such as trip characteristics (e.g. speed, acceleration, route, distance) for each trip can be selected in the mode detection algorithms.

The characteristics required in trip-based transport modelling are:

- geographic origin/destination,
- start and end time,
- mode,
- itinerary (route)
- purpose of a trip? GPS-based data lack information on the purpose of the trip.

Data on Points of Interest of the destination must be considered to improve the input characteristics. Geographical information on the land use of the destination has a significant correlation with the purpose of the trip. The addition of contextual data based on detailed geospatial information systems can significantly improve the expected results of the predictions (Sadeghian et al., 2021). This aspect is common not only to the prediction of the trip's purpose but also to the prediction of the daily activities carried out by a citizen in a place where he or she stays for a suitable period of time.

Unfortunately, geocoded land use databases often lack accuracy because they are not well maintained. Even if the data is up to date, it may not be possible to distinguish between travel purposes. For example, some locations may be classified as mixed use, i.e. having both commercial and residential uses. Data on different types of land use and points of interest (POIs) can be used to complement GPS to deduce transport modes such as train, light rail, bus and metro. The distance to stations and stops can be used to classify transport modes.

HETUS's daily activity tracking is also based on comprehensive data on existing land use and popular locations around travel destinations (Cheng et al., 2022). In order to extract features from the raw POI data, some issues need to be carefully considered: how to aggregate trivial POIs; how to encode POI data.

The approach outlined so far is therefore a hybrid one in which the statistical model adopted to predict survey variables combines both a dynamic model for extracting features from trajectories and a static model for extracting geographic data from maps. The model aims to overcome the accuracy limitations of ad hoc rule-based approaches, as such deterministic rules tend not to capture the stochastic nature of GPS data.

2.2.3. Accuracy of prediction and quality of input

Developing and validating predictive models, using ML approaches, to infer survey variables, in particular transport mode, from GPS data is a complex and difficult task.

The results of any approach will depend on:

- the accuracy of the measurements and statistics generated;
- the distance and speed from the GPS device;
- the absence of variability in the relevant GPS measurements between respondents, time and spatial context;
- the sensitivity of trip detection algorithms;
- the inherent differences between modes of transport in specific contexts;
- the ability of the inference system to detect these critical differences between modes of transport with respect to the selected variables, perhaps taking into account the temporal and spatial context;
- the quality of map services (contextual data).

The GPS data itself, and the subsequent algorithms used to segment the tracks and predict travel mode, trip purpose and activities undertaken at a location, are not fail-safe: GPS tracks may be incomplete, with missing data due to people's behaviour or the situation in which the GPS signal may be lost; GPS tracks may be complete, but, for example, the trip detection algorithms may be too sensitive (over-identifying trips) or not sensitive enough (under-identifying trips).

Heterogeneity in GPS sensor quality between different types of smartphones affects the measurements and the statistical information derived from them. This heterogeneity becomes a crucial issue in the context of the ESS, since an uneven distribution of smartphone brands and models can be observed across countries, as shown in Appendix A.

The improving contribution on the predicted variables of contextual data used to add location details may be limited by the quality of the map services themselves, which again may vary from one European country to another.

To assess the quality of map elements, one can use the requirements defined by ISO (2022) for geographic data, which are:

- Completeness - describes the presence or absence of features, their attributes and relationships.
- Logical consistency - refers to the degree of adherence to logical rules of data structure, attribution and relationships.
- Positional accuracy - measures the closeness of a feature’s position to values as accepted as or being true.
- Thematic accuracy - comprises the accuracy of quantitative attributes, the correctness of non-quantitative attributes and the correctness of the classification of features and their relationships.
- Temporal quality - describes the quality of temporal attributes and temporal relationships of features.
- Usability - is based on user requirements and assessed by evaluating quality along the other dimensions.

Data quality refers to the degree to which the requirements of a set of intrinsic data characteristics are met. Geospatial data are generally collected and used for many different purposes; the quality of geospatial data must be evaluated considering the purpose and context in which they are used.

An important quality indicator of geographic data for use in predictive modelling is the completeness of map elements (ISO 2022), in particular points of interest (POIs), as it directly influences the prediction of trip distribution, trip purpose and stops. A convenient approach is to examine the quality of different object classes, such as the road network. The most accurate method of quality assessment is to compare a dataset with its true value. However, estimating the true value is expensive, complex and time consuming. A more appropriate method is to compare the quality of a dataset relative to a benchmark dataset of documented high quality.

This issue has been explored in depth for the POIs contained in the map, comparing the two main mapping services used in the GPS-based methods, Google Places (GP) and OpenStreetMaps (OSM) through an experiment. This experiments been carried out randomly selecting points within urban areas of 11 countries: the study compares POI data detected within a 50-meter radius around each point. The study highlights a significant difference between the countries and a significant under-coverage of OSM compared to GP. The results are detailed in Appendix A.

2.2.4. The methodology implemented in the SSI Geotracking Microservice

In this section some methodological choices underlying the Geotracking microservice are presented and discussed. The Microservice is constructed as two distinct microservices. The first is a non-domain-specific microservice, while the second is domain-specific. The end-to-end process aims to assist users/respondents in completing their timelines through the microservice. The development of the microservice is overseen by WP3 and it is mainly described in WP3 deliverable 3.2.

We do not present the algorithm for GPS point segmentation, developed by WP3 and illustrated in deliverable 3.2. It processes GPS data, after a phase of pre-processing, producing stop and tracks with the addition of contextual data derived from the map service. These variables constitute the input data for the two successive phases, the transport mode prediction and the activity prediction.

2.2.4.1. Travel mode prediction

Transport mode classification is possible after the geolocation data is segmented into stops and tracks.

This approach requires a database with information about transport mode infrastructure, such as OSM. OSM was chosen because it is easily accessible and to not rely on a commercial platform/provider. One option to determine which transport mode is used in a track cluster is to map the geolocation data of the segmented track clusters to the infrastructure data.

Other approaches, such as machine learning, can be considered, but it was decided not to do so, as previous results were not promising (Smeets et al., 2019). After mapping the geolocation data to the OSM data, the number of OSM geolocation points per transport mode within a track cluster needs to be determined. The transport modes available in OSM are motorized vehicles on roads, trains, trams, buses, subways, bicycles, and on foot. It can be calculated which transport mode has the largest proportion in the track cluster considered. The transport mode with the largest proportion is then considered to be the most plausible mode and can be assigned to this cluster.

Depending on the infrastructure, it can be the case for certain track-clusters that no OSM is available for a transportation mode. For example, pedestrian paths are rarely available near highways. It is also possible to generate multi-modal clusters. This is the case when different transport modes have the same proportion in a cluster. In these scenarios, respondent interaction might be required to select the correct transport mode.

The quality of the infrastructure data is particularly important in this approach. If different transport modes have different numbers of data points, this can lead to biases towards or against certain transport modes. The quality and density of the various transportation modes can vary depending on the country. For example, the number of subway and tram location points in the Netherlands is relatively small, whereas the number of roads, cycling paths, and sidewalks is very large.

There are still a number of open questions at this step, such as the data quality and comparability of the infrastructure data, how to deal with multi-modal track clusters, but also how different segmentation algorithms affect this method. Furthermore, deterministic rule-based approaches taking into account, for example, the length of the track (distance and time) or the average speed of the track, have to be developed and implemented. Finally, quality criteria are missing at the current development stage and also need to be developed.

2.2.4.2. Activity prediction

The last part of the microservice associates HETUS activities distribution (with assigned probabilities or scores) to each stop identified in the first microservice.

This is performed through a model that exploits (following the reasoning in Cheng et al. 2022) several pieces of information, such as place categories taxonomy, timing of the stop, country-specific indicators derived from previous HETUS survey data in specific countries and user characteristics. Categories of place from the third party (Google Places GP or OSM) are mapped to the HETUS classification of places, to connect them to HETUS activities. The input data used for the activity prediction algorithm are:

- GPS points information:
    - Longitude
    - Latitude
    - Accuracy
    - Timestamp
- Stop attributes
    - Start time
    - End time
    - Duration
    - Centroid
- Map elements - Points of Interest (POI) inside the radius of the stop:
    - POI Longitude
    - POI Latitude
    - Place name, place category
- Profile of the user
    - Age class
    - Occupational status (employed, student, other)

For each stop identified and enriched with features derived from GPS and contextual data (map services and user characteristics) the algorithm can be summarised in the following steps:

1. POIs identification and selection of a short list of POIs in each stop.
    - A score (POI-score) is assigned to each POI inside an adaptive radius around the stop centre location, based on the weighted median of the distances calculated between each POI and all GPS points of the stop, weighting by the accuracy of GPS points.
    - A short list of POIs is identified using the elbow criterion on the POI scores.

2. Determination of (conditional) probability of HETUS activities for each POI selected in the short list.
    - Thought a Bayesian decomposition, for each POI in the short list the conditional probability of HETUS activities are calculated starting from the distribution observed in TUS data.
    - The variables considered (duration and time of the day, HETUS place category, occupational status, age classes) in the decomposition are linked with the corresponding variables observed in the stop and for the specific respondent.

3. Assignment to the stop of a rank of the HETUS activities based on a final score.
    - Final score is calculated aggregating the probabilities of the activity weighted by the POI-score associated with the activity for each POI in the short list.

The procedure has been tested so far on the data set collected using the CBS ODIN app in a version developed to be able to collect data in Italy. This phase is ongoing at the time of writing and will consist of validating the resulting activity prediction by comparing it with the actual activities performed by users (recorded via an annotation). An additional data set of GPS data will be collected to test the algorithm on a wider set of respondents.

The comparison will also be made between the predictions obtained using the OSM and GP POIs.

The main steps of the activity prediction procedure are displayed for an example stop derived from the mentioned data set in the following figure. The first map represents a stop and contains the GPS points (in green) and the POIs on the map (in blue); the second map highlights in red the POIs on the short list and the final table shows the ranking of activities with corresponding weighted scores.

HETUS	ActivityScore	Descr
021	7.400402e-02	021 Eating
361	4.739460e-02	361 Shopping (including online/ e -sho
519	3.842740e-02	519 Other or unspecified social life
032	3.437884e-03	032 Personal care servi ces
732	1.588585e-03	732 Parlour games and play
513	1.237589e-03	513 Celebrations
821	1.227424e-03	821 Watching TV, video or DVD
522	3.925254e-04	522 Theatre and concerts
343	3.686005e-04	343 Caring for pets
831	3.219861e-04	831 Listening to radio or recordings
383	1.547464e-04	383 Reading, playing and talking with c
811	8.229526e-05	811 Reading periodicals

Figure 2.1: Main steps for the activity prediction, input and output

2.2.4.4. Activity prediction assessment and test

The presented Geotracking microservice is under development in the SSI project: the main pipeline steps have been implemented but the assessing and testing phase is ongoing.

Quality assessment is a crucial part of the work of the SSI project in the coming months, to evaluate the impact that different choices and different types of input data have on the goodness of the predicted variables, which will then be used as tentative data in the UI, through the interaction with the respondent.

In order to assess the quality of the activity prediction algorithm, it is necessary to perform an impact analysis, trying to answer several questions.

- How much the prediction depends on the use of TUS data and the user profile?
- How sensor and map quality influence the quality of prediction?

In the next steps, it is necessary to carry out tests and output evaluations, but also to test:

- The implementation of the microservice in different platforms.
- The microservice in large and small test in different countries.

2.2.4.5. Country-specific issues

In the context of GPS data processing, country specificity issues are more mitigated than in the context of HBS. In fact, what is linked to the country is above all the map, both in terms of the possibility of using the API to Google Places and for the quality and updating of the information contained in the map service used.

The phenomena investigated through the use of GPS data, movements and daily activities, evolve slowly and do not present a problem of constant updating except for the map, which however, if it is GP, is constantly updated by the third party.

In the geotracking case the country specificities have no impact on the algorithm, but only on the quality of the output in relation to the quality of the input (sensor and GPS quality) and of the auxiliary information. The issue of the specificity of the model parameters (for example for the use of indicators estimated from the TUS survey) is perhaps the only one.

2.3 Conclusions: towards the final deliverable on methodological guidelines for ML

The remaining months until the end of the project will be used to conclude the implementation of the two microservices and to evaluate their performance, both with internal and field tests.

The next steps for the AI-ML in receipt processing are:

Receipt text extraction:

- Evaluation of convergence of the text box classification accuracy and its dynamics in time
- Guidelines on the collection of annotated train data

COICOP classification:

- Selection of the string matching distance
- A choice of ML method(s) for COICOP classification including the inclusion of hierarchy
- Refinement of the three-step procedure including specific thresholds and guidelines
- Guidelines on retraining of ML classification

For the geolocation data, strictly speaking there is no AI/ML involved yet since the prediction models are based, for the moment, on the use of contextual data. It is likely that there will no AI/ML involved for transport mode prediction nor for activity prediction.

In this context, the next steps consist in the assessment outlined in section 2.2.4.4, which will serve both to complete the development and release of the methodology for the geotracking microservice, and to bring out and outline indications and guidelines on the processing of GPS data in smart surveys.

References:

Sadeghian P., Håkansson J., Zhao Xiaoyun. (2021). Review and evaluation of methods in transport mode detection based on GPS tracking data. Journal of Traffic and Transportation Engineering (English Edition) · July 2021. DOI: https://doi.org/10.1016/j.jtte.2021.04.004

Cheng J., Zhang X., Luo P., Huang J., Huang J. (2022). An unsupervised approach for semantic place annotation of trajectories based on the prior probability. Information Sciences, Volume 607, August 2022, Pages 1311-1327. Elsevier ed. https://doi.org/10.1016/j.ins.2022.06.034

ISO 2022. Geographic information — Data product specifications

3. Human Computer Interaction and Usability: insights from workpackage 2.3

3.1 Introduction

The aim of the SSI project is to develop, implement and demonstrate the concept of Trusted Smart Surveys, realising a proof of concept for the complete, end-to-end, data collection process and demonstrating a solution. This solution combines:

1. The involvement and engagement from citizens.
2. The acquisition, processing and combining of data collected from smart devices and other applications.
3. The contribution to trustworthiness and guarantee of strong privacy safeguards.

As outlined in the SSI project's M6 deliverable, the usability of the solution is the key concern within the Human Computer Interaction (HCI) spectrum. Usability refers to the ease of use and the quality of the user's experience with a platform or application. It has been argued that usability is strongly linked to the three key elements of the solution provided within the SSI project:

Ad. 1. The involvement and engagement from citizens relate to the usability of the application to recruit and retain users. Key usability attributes are engagement, accessibility, clear instructions, time efficiency, and the availability of error handling options.

Ad. 2. The acquisition, processing and combining of data collected from smart devices and other applications relates to the usability of the application to complete complex tasks. Key usability attributes are a clear and intuitive user interface (UI), a clear task flow and guidance, error prevention mechanisms, (in app) training, and (in app) feedback and support.

Ad. 3. The trustworthiness and guarantee of strong privacy safeguards relate to the usability of the application to share personal data. Key usability aspects are trust and credibility, security and privacy, transparent communication, data collection efficiency, and user control of data/information.

The goals set out in the SSI project focus on (a) an end-to-end solution for a smart survey framework that is sufficiently mature to be applied in several ESS applications, (b) 'going smart' by conceptualizing and operationalizing several micro services, (c) identifying developing, implementing and eventually demonstrating these microservices within the perspective of an end-to-end process for (at least) Time Use Surveys (TUS) and Household Budget Surveys (HBS), and (d) considering smart surveys from a respondent perspective.

The usability testing is mainly concerned with the latter, albeit some tests may relate to the design and implementation process of the microservices (cf. c). This section outlines the protocol for usability testing of the receipt scanning microservice as part of the HBS. The Think Aloud (TA) protocol has been identified as the most relevant methodology to conduct this test (see M6 deliverable). The TA user testing method consists of users conducting different tasks while verbally articulating the struggles or experienced difficulties when doing a task and is aimed at investigating the *usability* of the application. Note that *usability tests* are a special form of user tests that focus on the ease of use and effectiveness of a system and typically involve scenarios or tasks that are designed to evaluate how easily users can accomplish these with the system.

3.2 Think Aloud Procedures

3.2.1 Set up

The VUB, CBS and Destatis will conduct the user tests focusing at User Interaction and Usability at a physical location. SSB will conduct the experiments online. The tests will take about 1 up to 1.5 hours excluding the participants' travel time. Roughly the test is subdivided into 10 minutes introduction, 25-35 minutes doing the tasks, 25 minutes debriefing. Participants are requested to use their own smartphone on which they have to install the applications (in case of MOTUS and @HBS; see further down) and bring several personal receipts for the test (for tasks 6 and 7; see further down). The observers will provide receipts for the other tasks (for tasks 3-5 and 8; see further down). Observers will provide a back-up phone to anticipate outdated operating systems or broken cameras as well as chargers to anticipate low battery levels on the participant's phone. Participants will be given five tasks to complete, while simultaneously verbalizing their thoughts (and not only their actions). There will be no active interaction between the participant and the observer. Audio and video (screen) will be recorded. Note that video recordings concern screen and hand/finger movements only. The face of the participant will not be recorded.

To conduct the tasks, participants need to be connected to the internet. At the physical locations, participants will be asked to connect to the guest WiFi and turn on flight mode and not disturb mode to prevent accidental recording of notifications of personal messages. After completion of the last task, the participant participates in a debriefing. At the end of the test, the observers will help the participant delete the application and all data as well as disconnect from and 'forget' the guest WiFi.

3.2.2 Platforms

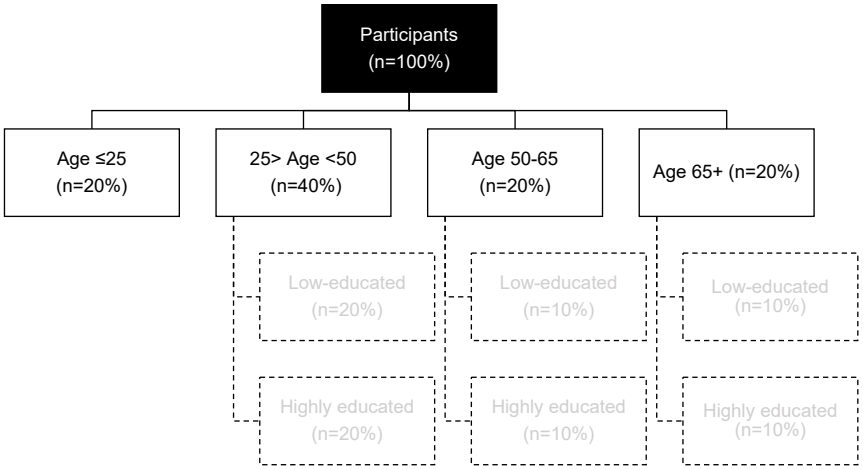
The TA tests for the receipt scanning microservice will be conducted with three different platforms. The MOTUS platform will be used by VUB and Destatis. MOTUS is developed by hbits. The @HBS app will be used by CBS. Both the MOTUS app and the @HBS app use the OCR microservice and classification microservice developed within this SSI project. Both apps need to be downloaded from the App Store or Google Play Store. The Progressive Web App (PWA) will be used by SSB. It can be opened in any net browser. It is not an app that needs to be downloaded from the App Store or Google Play Store. In all applications, participants can scan tickets and edit information retrieved from the ticket.

3.2.3 Participant recruitment

The case demonstrations from the TA tests of the end-to-end smart survey solutions (here: HBS) are targeted at population samples. For the usability testing method, it is recommended to recruit participants from these populations. Overall, it is decided to keep a wide focus and preferably stratify recruitment along the dimensions of age, and – if possible – education (see Figure 3.1). Additionally, we aim to (also) include participants who do not use ICT for their work. Since sample sizes might differ somewhat by country and subgroup size, the sample size in Figure 3.1 is presented as a percentage of the total sample.

The recruitment of participants can possibly take place via existing channels of the National Statistical Institutes (NSIs) such as panels, via networks of civil society organizations, or calls on social media.

Figure 3.1 Participants' characteristics



3.3 Study design

The participants will be asked to complete six tasks in a designated order. The first task is a warming up task or primer for the participant to familiarize themselves with verbalizing thoughts. For the third task, the participant is presented with a short, clean receipt and asked to enter the expenses in the expense diary. The participant is free to choose between the 'scanning' and 'manual' method. For the fourth task, the participant is again presented with a short, clean receipt, but now the participant is forced to use the method not chosen in the second task. For the fifth task, the participant is presented with a receipt that is known to require editing because of lower quality and the participant is forced to use the 'scanning' method. If the participant does not edit the receipt, task 5-bis will request the participant to edit. For the sixth task, the participant uses a personal ticket but is free to choose the method. In case of scanning, if the participant does not edit the receipt, task 6-bis will request the participant to edit. The seventh and/or eighth task are optional and depending on the time. If the participant used the 'manual' method for task six, the next task is task seven. The participant again uses a personal ticket but is forced to use the 'scanning' method. If the participant used the 'scanning' method for task six, the next task is task eight. The participant is presented with a long receipt that needs editing, and the participant is forced to use the 'scanning' approach. In both cases, if the participant does not edit the receipt, task 7-bis and 8-bis will request the participant to edit. The eighth task is platform specific and requests the participant to consult the expense overview. Table 3.1 gives an overview of the tasks.

Table 3.1 Task overview

Task #	Receipt type	Usability focus	Rationale	Observer focus
Task 1	-	Warm-up/primer	Practice think-aloud	Actively stimulate verbalizations of thoughts
Task 2	-	Install app and log-in	Prepare/access	
Task 3	A1. Short and clear	Choose approach	Start easy	

Task 4	A2. Short and clear	Force inversed approach	Make sure both approaches are tested	
Task 5	B1. Medium in need of editing	Forced scanning approach	Continue with more complex task/ticket, that we know needs editing.	Does the participant notice errors from scanning and does the participant edit errors?
Task 5-bis		Forced editing	A chance to edit errors.	
Task 6	C1. Personal	Choose approach	Continue with personal ticket. Which approach they choose now that we know that they are aware of both approaches.	Are considerations about sensitivity and personal information more prominent when sharing their own receipt?
Task 6-bis		Forced editing	A chance to edit errors.	Do they notice errors from scanning? Do they edit errors?
Optional, depending on time.				
If used 'manual' method for task 6.				
Task 7	C2. Personal	Forced scanning approach	Continue with personal ticket. Which approach they choose now that we know that they are aware of both approaches.	Are considerations about sensitivity and personal information more prominent when sharing their own receipt?
Task 7-bis		Forced editing	A chance to edit errors.	Do they notice errors from scanning? Do they edit errors?
If used 'scanning method for task 6.				
Task 8	B2. Long in need of editing	Forced scanning approach	Continue with more complex task/ticket, that we know needs editing.	Does the participant notice errors from scanning and does the participant edit errors?
Task 8-bis		Forced editing	A chance to edit errors.	
Platform specific				
Task 9	All tickets	Find and check Expense Overview	Navigation and usability	

3.3.1 Explaining the test

Before starting, the observer welcomes the participants, explains the test and requests consent. Note that the test can only start after the consent form has been signed. The explanation should at least address the following elements:

- How long the test will take.
- The aim of the test.
- What is expected from the participant (i.e., verbalizing thoughts when doing tasks).
- What will happen (i.e., audio and video recording tasks done using the app).
- Confidentiality of participation and option to ask any questions on pre-distributed Privacy and Confidentiality agreement.

- Option to ask any questions on (pre-distributed) consent form and signing off the consent form.
- Option to ask any other questions.

Countries use their own consent form. The VUB needed approval from the university's ethical committee to conduct the tests, for which the application has been submitted on April 10 and which has been granted on April 29 (ref: ECHW_510).

3.3.2 Tasks explanation

The observer reads out the tasks one at a time. Observers should not give any extra information or direction about how the tasks need to be completed. See further down for instructions on observer-participant interactions. *Note:* Participants need to have WiFi or Data access to complete the tasks.

Once recording started, participants start with the first task, which serves as a warm-up or primer task to practice thinking aloud. The task involves finding a route between two points (to be provided by the observer) using a navigation app such as Google or Apple maps. The observer may explicitly prompt and/or explain how to verbalize their thoughts.

In their second task, participants should use an excerpt from the information in the invitation letter to find the application in the App/Play Store, install the application on their smartphone, consent to the app requirements, and login to the application. *Note:* this task is marked as finished when the participant is at the home screen of the application. Allow some time for participants to explore the home screen before presenting the next task. If they do, the observer will interrupt participants after one or two minutes and suggest proceeding to the next task.

In their third task, participants need to add expenses from the receipt to the HBS expense diary. **Observers are instructed to ot give instructions to scan the receipt**. The participants are free to choose between the 'manual' and 'scanning' method. Depending on the way participants added the previous expenses to the HBS expense diary, their fourth task is to use the other method of adding expenses. This will either be a) using the manual input or b) using the receipt scanning option. **Observers should not tell participants how to access the receipt scanning.**

In their fifth task, participants are forced to use the 'scanning' method but are given a receipt that needs editing. **Observers do not initially tell the participant that the receipt needs editing.** However, if the participant does not edit the expenses, force the participant to edit the expenses (i.e., task 5bis).

In their sixth task, participants use a personal ticket and are free to choose the method. Knowing both methods and having practiced with the editing after scanning, this allows to assess participants' preference and tells us something about the usability of both methods. It also allows observing participants' considerations about privacy and sharing personal data and, hence, the trustworthiness of the application.

Depending on time, the next task continues with more forced scanning, either using another personal receipt if chosen the 'manual' method for task six (i.e. continue with task 7) or using a long receipt if chosen the 'scanning' method for task six (i.e., continue with task 8). Additional 'scanning' tasks provide more insights into the usability of the OCR microservice.

Finally, task nine is platform specific and requests participants to find the overview of the expenses that they have added to the expense diary.

The wording of each of the tasks is listed in Table 3.2.

Table 3.2 Overview task wordings

Task #	Wording	Material	Notes
Task 1	Please use an app of your choice to find a route between XXX and XXX.		
Task 2	This is the invitation letter that invites people to participate in the study and install the app. Please read the letter and proceed as described.	HBS invitation letter	*The participants use a pre-set account with username and password as listed in the letter. The participants do not need to create an account, provide an email address, and set a password.*
Task 3	Imagine you have been shopping at {{store name}} and this is your receipt. Please add these expenses in the expense diary in the application.	Receipt A1.	
Task 4	Imagine you now have been shopping at {{store name}} and this is your receipt. Previously you used the **receipt scanning/manual entry** method to add the receipt to the expense diary. Please add these next expenses **manually/by scanning the receipt** to the expense diary.	Receipt A2.	*Force other method than used in task 3.*
Task 5	You have done some more shopping, this time at {{store name}} and this is your receipt. Please add these next expenses to the expense diary using the receipt scanning option.	Receipt B1.	
Task 5bis	You scanned the receipt to add it to the expense diary. Please locate the expenses, check and if necessary, edit them and commit the expenses to the expense diary.		*Force editing.*
Task 6	Now let us try to add the expenses of your own receipt. Please choose one of your receipts and add these next expenses to the expense diary using the method of your choice.	Receipt C1.	*Free choice of method.*
Task 6bis	You scanned the receipt to add it to the expense diary. Please locate the expenses, check and if necessary, edit them and commit the expenses to the expense diary.		*Force editing.*
Optional, depending on time			
If used 'manual' for task 6.			
Task 7	Previously you used the **manual entry** method to add your own receipt to the expense diary. Please choose another of your receipts and add these next expenses **by scanning the receipt** to the expense diary.	Receipt C2.	*Force scanning.*
Task 7bis	You scanned the receipt to add it to the expense diary. Please locate the expenses, check and if necessary, edit them and commit the expenses to the expense diary.		*Force editing.*
If used 'scanning' for task 6.			
Task 8	This is your last task. You have one more receipt. This time from {{store name}}. Please add these next expenses to the expense diary using the receipt scanning option.	Receipt B2.	*Force scanning*
Task 8bis	You scanned the receipt to add it to the expense diary. Please locate the expenses, check and if necessary, edit them and commit the expenses to the expense diary.		*Force editing.*
Platform specific.			
Task 9	Please try to get an overview of all the expenses you have entered.		

3.3.3 Tickets

In light of this SSI project, the core of the test is the usability to enter, scan, and edit receipt information to be committed to the expense diary and the participants willingness/engagement with and trust in the application/solution. To be able to compare results/discussions across countries, it is important that tasks and tickets are highly comparable across countries. Table 3.3 lists the ticket characteristics per task.

Table 3.3 Ticket characteristics

Task #	Receipt #	Characteristics
Task 1	-	-
Task 2	-	-
Task 3	A1.	- Grocery store receipt - Short (3 products) - Clear (good quality print) - Recognizable product abbreviations
Task 4	A2.	- Grocery store receipt - Short (3 products) - Clear (good quality print) - Recognizable product abbreviations
Task 5	B1.	- Grocery store receipt - Medium (8-10 products) - Unclear (medium quality print) - Recognizable product abbreviations
Task 6	C1.	- Personal ticket of respondent
Task 7 (if task 6 'manual')	C2.	- Personal ticket of respondent
Task 8 (if task 6 'scanning')	B2.	- Grocery store receipt - Long (15-20 products) - Clear (good quality print) - Recognizable product abbreviations
Task 9	-	-

3.3.4 Debriefing

After the participant has finished the last task, there is a short debriefing during which the participant is asked to answer a few more questions first about the tasks and second about their perceived trustworthiness, credibility and security of the application in general and receipt scanning microservice in particular.

3.3.5 Observer roles

Throughout the test, the observer takes on different roles. The roles and accompanying (inter)actions are described below. The overall aim is to stick to these roles as closely as possible to keep observer roles standardized across the different tests.

Setting the stage. Observers need to explain the purpose and set up of the test. They need to make clear that the test is about the usability of the application the participant will use to complete the tasks. It is not about the participant or about the participant's ability to complete the tasks. The observers clearly explain that the participant is the 'expert'; the important contributor and that the observer is there to learn and listen. Section **Error! Reference source not found.** contains an introduction text.

Additionally, the observer needs to obtain consent from the participant to take part in the test, to be audio and/or video recorded, and to use an application that will store their data. Section **Error! Reference source not found.** contains the consent request.

Active listening. Observers need to learn and listen and will do so following the speech communication protocol. This allows observers to listen actively by using acknowledgement tokens. Only use the tokens agreed upon in this protocol:

1. Non-verbal: head nodding
2. Back-channelling: noncommittal responses such as “mhmm” / “hmmm” / “right”

Observers should avoid verbal affirmations of agreement such as “Yes” / “Yeah” / “Yep” / “Uh-huh”, disagreement such as “No” / “Nope” / “Nah” or understanding such as “I see” / “I understand”.

Additionally, the observer needs to keep the participant talking by using the following acknowledgement tokens only:

- Encouragements: “go on” / “continue” / “and then?”

In terms of frequency of use, the acknowledgement tokens should follow the flow of current communication. The observer should avoid proactive interventions such as asking for clarifications or probing for more information. Make notes and come back to your notes in the de-briefing.

Interactions. At different points in the tests, interaction might be needed. Observers must try to follow the protocol for different interactions as strict as possible.

- Instructions: Hand out the needed material and read out the instructions.
- Technical issues: In case the participants face technical issues, try to fix the problem silently. Any comments while fixing issues might influence the participant's approach to the task. If technical issues are substantial, explicitly interrupt the test.
- The participant thinks the task is completed when it is not: Remind the participant of the task and ask the participant to continue without giving direction for the solution.
- The participant is stuck: Remind the respondent to try and solve the task without help, as if you are not there (i.e., which would be the case in a real survey). If the participant is still stuck, use questions like “What are you trying to do?”, “What do you expect to happen?”, and “What are you looking for?” to understand where the participant gets stuck. If the answer is something the application can do, point the participant in that direction. If the answer is something the application cannot do, suggest another direction towards the solution. Remember, assist only if necessary to move the task forward. Reaffirm the participant that it is not about their ability, but the applications usability to allow task completion and that this is an as relevant result as completing the task.

Note: If the participant gets stuck at a crucial step in the workflow, point out the “solution” to the participant in order to be able to run the other tasks.

- The participant asks a question about the task: Explain the task again without giving direction for the solution to complete the task.

Note: if needed, explain to the participant that the aim is to see how they would deal with the app and the tasks in a “normal” situation (i.e., with the interviewer not being present). Note down the question to get back to it during the task debriefing.

- The participant starts a “meta discussion”. For example, asking the observers’ opinion, whether the observer tried to do it him/herself, that it will never work, that they do not understand that scientists really think people will use this app or feature, et cetera. Please respond using the standardized sentence: “I understand your remark/question/concern and I consider it very valuable. For the test to be successful, I do not want to influence you too much and would like to ask you to continue with the task. Afterwards we can discuss your valuable feedback further.

In general: The essence of the usability test is not to engage in conversation with the participants. Observers need to probe participants to continue verbalizing their thoughts. However, if, and only if, the observer feels that participants are withholding crucial thoughts, the observer can ask the participants how they think the task went. **It is important that observers only acknowledge the participants’ comments** and to keep an eye on the timing. If participants have questions then the observers are instructed to tell participants that they will get back to their thoughts once all tasks are completed.

3.4 Analysis plan

The TA test will generate the following data sources:

1. Video (with audio) recordings
2. Observer notes
    a. Context (setting description)
    b. Field notes (i.e., questions of participants, scorings of task performance, etc.)
3. Debriefing questions
4. Performance measures (if applicable)

The tasks that the participants are requested to complete are designed in a way that they touch upon the different usability attributes that relate the three elements of the solution to a smart survey implementation (see Introduction). The data sources will be analysed and triangulated by all partners to complete a grid that relates the positive and negative, cognitive and emotional, and task-related verbalizations of the participants to the relevant usability attributes and overarching elements for each task cross referenced to the specific actions that are part of the workflow to complete the task (e.g., selecting camera, editing data, committing data to expense diary, et cetera). A simplified structure of the analysis grid or data chart at the partner level is included in Table 3.4.

Table 3.4 Analysis grid

Elements	Involvement and engagement from citizens.	Acquisition, processing and combining of data collected from smart devices and other applications.	Contribution to trustworthiness and guarantee of strong privacy safeguards.	Other comments.
Usability attributes	• engagement • accessibility • clear instructions • time efficient, • the availability of error handling options	• clear and intuitive user interface (UI) • clear task flow and guidance • error prevention mechanisms • (in app) training • (in app) feedback and support	• trust and credibility • security and privacy • transparent communication • data collection efficiency • user control of data/information	n/a
Task 1	n/a	n/a	n/a	n/a
Task 2	main focus			
Task 3		main focus		
Task 4		main focus		
Task 5		main focus		
Task 6		main focus	main focus	
Task 7 (if task 6 'manual')		main focus	main focus	
Task 8 (if task 6 'scanning')		main focus		
Task 9		main focus		
Trust debriefing			main focus	

3.5 Conclusion

This chapter outlines the protocol for the small-scale tests on microservice for the Household Budget Surveys (HBS) within the SSI project. The main aim of the small-scale tests is to assess the ease of use and effectiveness of the receipt scanning microservice for HBS. This microservice – in combination with a classification microservice – should make it possible to register expenses in the household budget diary in a faster, less burdensome and smarter way. By tasking test persons with different types of receipts and different registration options (i.e., manual and scanning), conclusions will be drawn on different usability attributes that relate to the proposed smart solution of receipt scanning, and to the acquisition, processing and combining of data collected from smart devices and other applications in general.

4. Mode effects and other data integration considerations: Insights from Work Package 2.4

4.1 Background

National Statistical Institutes (NSIs) can benefit from the introduction of smart features to surveys, increasing the quality of the resulting data by making the data collection process easier for participants, or by offloading certain tasks that are better handled by sensors or machine intelligence. A side effect of data quality improvement is that the data generated by a smart survey may differ meaningfully both in representation and measurement. While this may not matter for any one cross-sectional study, official statistics are generally subject to rigorous requirements for similarity over time to be able to demonstrate patterns over a longer period. To this end, surveys are often constrained to be as similar as possible with respect to measurement differences, and complex models are developed to account for changes in representation from year to year. The objective of work package 2.4 (WP 2.4) is to evaluate the differences that smart features introduce, and to propose methods for integrating smart and traditional surveys.

4.2 Challenges for integration

Deliverable M6 identified two fundamental questions:

> *1. Where there is measurement error in an existing survey that may be corrected in the transition to a smart survey, how should this be considered with respect to the integration of the data sources?*

Measurement error in surveys can stem from various sources, either from the respondent, such as with social desirability bias and satisficing bias, or the survey instrument itself, impacting its usability. Because smart surveys will differ from non-smart surveys in the interaction between these elements, we can expect this to contribute to the overall mode effect in a meaningful way. For example, in Time Use Diaries (TUDs), respondents tend to overestimate certain tasks and underestimate others (Kelly et al., 2015; Harms et al., 2019; Sullivan et al., 2020), and the intricate categorization process is often tricky for users to navigate (Bauman et al., 2019). By incorporating smart features such as geolocation measurement, researchers can alleviate the burden on respondents by completing some fields and providing additional data to improve recall. The integration of a search feature for the categorization of an activity can lead to a more accurate categorization process than the navigation of a hierarchical tree structure. On the other hand, both additions are intentionally designed to elicit differences in measurement between the non-smart survey and the smart survey.

All smart surveys are designed to offload some portion of the response generation process onto the user; to this end, researchers hope to reduce careless reporting (Conrad et al., 2005), missing data (Chatzitheochari et al., 2018; Lev-On & Lowenstein-Barkai, 2019), inaccuracies due to either to misunderstanding (Schwarz, 2012), or lack of knowledge (Greaves et al., 2015; Gillis et al., 2023), or social desirability bias (Keusch, Bach, et al., 2022). In some cases, smart surveys may also increase measurement error relative to paper diaries; they may generate less detailed responses (Frąckowiak et al., 2022), or select app-provided defaults more frequently (Bucher Della Torre et al., 2017). Any smart survey that must be compared with a non-smart survey must consider avenues of potential measurement differences in the survey's design.

2. If representation differs between smart surveys and traditional surveys, how can this be disentangled from measurement differences?

The literature review in the M6 deliverable indicated that differences in representation were likely to occur between smart- and non-smart surveys. Chapter 1 of this deliverable provides a more thorough discussion of this topic, identifying two categories of factors that can influence the decision to participate: characteristics that researchers can control directly (e.g., invitation design), and those that they cannot (e.g., personal characteristics of the sampled person). While good design can remove barriers related to the first category, some obstacles related to personal characteristics will remain. For example, persons with no smart device cannot participate in an app-based study, and the smart perceptions survey from Deliverable 1.2 revealed smartphone ownership rates of only 86.1%, 88.1%, and 91.5% across the three surveyed countries (Lunardelli et al., 2024, Table G.5). Because smartphone ownership remains unevenly distributed in the population, offering alternative non-smart modes remains critical for reducing under-coverage (Klingwort & Schnell, 2020; Keusch et al., 2023), despite increasing analysis complexity.

Where researchers have used smart surveys for Household Budget Survey (HBS) studies, analysis has indicated a combined selection and measurement difference (Riegler, 2015; Jäckle et al., 2019.) It may be possible to use participant characteristics to distinguish between measurement differences and representation differences (Premkumar et al., 2023), but this approach is not without problems (Tourangeau, 2017). When feasible, mechanisms for distinguishing between the two such as random assignment to mode condition, either with or without repeated measurement, can be integrated into the survey design (Schouten et al., 2013; Klausch et al., 2015.) The cost of gaining this precision is a reduced proportion of the sample benefiting from the smart survey mode. A key component of answering the second question is determining the extent of this interaction and thereby determining whether continued random assignment to non-smart modes is justified or even necessary.

The resolution of these two concerns, how we consider measurement differences between modes, and how we distinguish these from representational differences will provide insight into how integration should proceed. The M6 Deliverable proposed two broad directions for integrating smart and non-smart data: 1) constrain the smart survey data, applying a set of algorithms to get the data into the same shape, or 2) maintain the smart survey data, calibrating the two sources against each other. In the first, integration and analysis proceeds similarly to traditional mixed mode design, and data are combined before they are analyzed. In the second, the data may be too different to be integrated into a single data set, so the estimates produced within each must be combined in some other way.

4.3 Fieldwork plans

France, Belgium, and Italy will conduct experiments within the field tests of their smart surveys to address these questions. To investigate differences in measurement, both Belgium and France will develop a paper-based (PAPI) diary to compared against an app-based diary, while Italy will compare two versions of an app-based diary with different levels of smart features. To disentangle representation and measurement, Belgium will vary the follow-up moment at which the PAPI alternative is offered, France will collect both PAPI and app-based diaries (offered sequentially) from all respondents within the TUS experiment, and Italy will unify the invitation across both groups by using an interviewer to help install the app. Table 4.1 provides a quick overview of the tests, smart features and research questions across countries.

Table 4.1 Comparison of planned mode effect field tests across countries

Field test	Methodology	Smart features	Research questions	IC
France TUS	Sequential/crossover mixed-mode design PAPI and CAWI (device non-specific)	a, b, c, d, e, g	Data quantity CAWI vs PAPI	1
			Data precision CAWI vs PAPI	1
			Interaction of non-response and measurement mode effects on quantity and precision	2
France HBS	Random assignment to PAPI-only condition or choice condition (PAPI/app)	a, b, f, g	How does mode of administration impact measures of household consumption?	1
			How do representational differences manifest between participants choosing the paper mode and those assigned to it?	2
Belgium TUS	Invitation to app, followed up by random assignment to PAPI alternative or no PAPI alternative.	a, b, c, e, g	Can data from PAPI follow-up responders be integrated in a mixed-mode manner?	2
			How does mode of administration impact measures of time use?	1
Italy TUS	Random sample split across high-smart and low-smart conditions in an app-based diary.	a, b, c, e, g, h, i, j	Does prepopulating the diary using geolocations impact the content of the TUS diary?	1
			Does prepopulating the diary using geolocations impact measures of quality of the TUS diary?	1

Note: **Integration Concerns** (**IC**) 1) Correcting for measurement mode effect, 2) Disentangling measurement and representation;
Smart features: a) field search, b) soft consistency checks, c) hard consistency checks, d) pictorial representations, e) time entry fields, f) image capture, g) auto-complete, h) auto-classification, i) geolocation, j) geoprocessing

The impact of mode on data quality is of interest for many countries. For both the HBS and TUS field tests, data quality impacts can be broken down into two categories: quantity, and precision. Measures of quantity are often based either on missing data itself, or on variables that may be used as a proxy for determining missing data: the number of distinct different categories that are recorded, the length of the day recorded in TUS, the total number of entries over a period of time in HBS, etc. Measures of precision are concerned with reducing inaccuracies in data that are recorded, and may involve comparison of start and end times, specificity of tasks, or whether diaries were filled in at one time, or spread throughout the day.

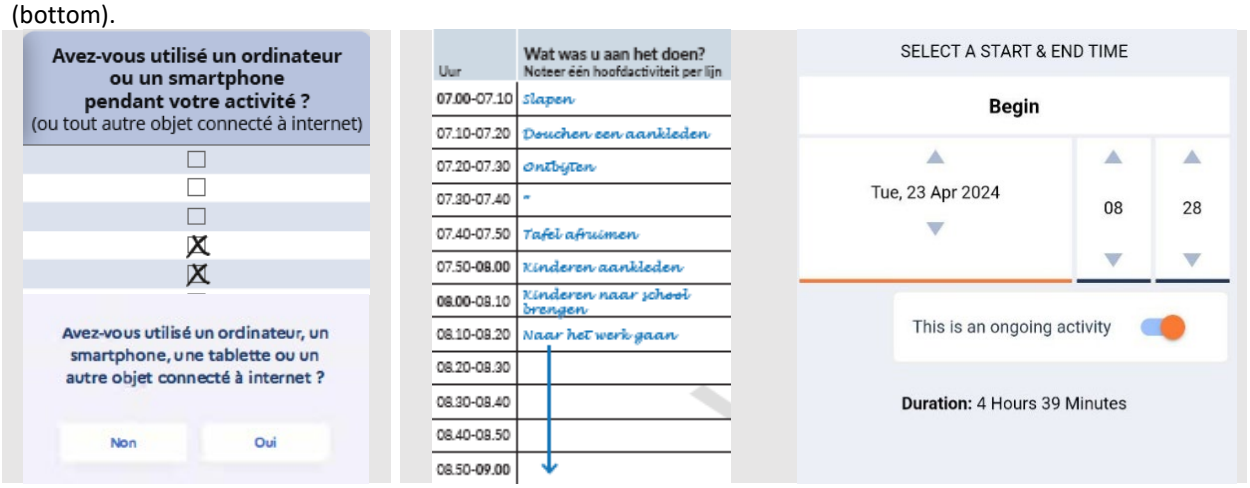
For data quantity, the expectation is that the addition of certain smart features, such as soft and hard checks, image capture, and geolocation will improve both visible (e.g. missing mode of transportation for a trip), and invisible (e.g. missing purchase) missing data. Figure 4.1 shows two comparisons of the same questions implemented differently across PAPI and app due to the integration of smart features. In subfigure a), a checkbox in the PAPI version asking whether or not an internet-connected device was used during an activity is replaced by a prompted question in the app, where a person must specifically indicate yes or no to the question, making missing data more visible.

For data precision, smart features such as time-entry, field search, and auto-classification are expected to demonstrate mode effects relative to the PAPI mode. Figure 4.1, subfigure b) demonstrates the comparison between the activity length designation. In the PAPI version, pre-printed ten-minute intervals are present in the diary, and respondents are asked to draw an arrow to indicate when a task begins or ends. In the app, this is replaced by a time entry widget, which may discourage users from extending task length incorrectly with the arrow, thereby omitting activities.

Figure 4.1: Example smart features hypothesized to improve data quality

a) Soft check smart feature, replacing PAPI checkbox (top), with prompted question in app (bottom).

b) Time-entry smart feature, replacing pre-printed 10-minute intervals in PAPI (left) with time entry widget in app (right).



A third category of analyses will involve calculating a set of the eventual analyses of interest to applied researchers and comparing the results between modes. For example, determining the distribution of childcare responsibilities or commuting time within a household.

France

Time-Use Survey

Methodology: Address-based household-based probability sample drawn from population register. Invitation via postal mail. A sequential mixed-mode design with randomized order assignment between the web-based (CAWI) and PAPI TUS diaries. The respondent is randomly assigned a day of the week (e.g. “Wednesday” or “Saturday”) and asked to complete the diary with respect to this specific day, reporting on the same day for both modes, but separated by one week’s time.

Smart features: Auto-complete search field for activities, (soft) consistency checks on incomplete diaries, (hard) consistency checks for some fields (date, activity versus travel), pictorial representations of hierarchical categories, time entry fields

Data quality measures: Interviewer's survey (interviewer’s rating of quality of data, whether/how many corrections were necessary); CAPI evaluation questionnaire (respondent reporting having filled in the CAWI mode in one go); diary content (length of day covered, time spent sleeping, number of meals, number of unique activities).

Household Budget Survey
Methodology: Address-based household-based probability sample drawn from population register. Invitations via postal mail, with random assignment to PAPI-only condition or choice condition (between PAPI and app-based)

Smart Features: Field search for activities, (soft) consistency checks on incomplete diaries, Image capture of receipts (without in-app processing).

Belgium

Time-Use Survey
Methodology: 8,000 Belgian households, proportionally selected in all three Belgian regions were invited via letter to participate in a 7-day app-based diary. Non-responders were subsequently randomly assigned to one of two conditions: a follow-up letter with a PAPI alternative or the same follow-up letter with no PAPI alternative.

Smart features: Field search for activities, (soft) consistency checks, (hard) consistency checks on missing fields, time and date entry fields

Data quality measures: Diary content (length of day covered, time spent sleeping, number of meals, number of unique activities, number of complete diaries)

Italy

Time-Use Survey
Methodology: Random sample split across high-smart and low-smart conditions in an app-based diary.

Smart features: Geolocations for segmentation of the day, visual prompting with maps, mode detection

Data quality measures: Diary content (Number of unique activities, start and end times, amount and content of missing data, specificity of tasks)

4.4 Analysis plans

France TUS

Research question(s):

- Is the amount of information provided in the CAWI diary equal to that of the paper diary?
- Is the precision/quality of the information provided by the CAWI diary equal to that of the paper diary?
- How can the navigational path the respondent uses to select an activity lead to differences in activity selection between modes?
- Is there an interaction between non-response and mode measurement effect on quantity and precision of data?

Analyses:

- Estimate overall mode effects for quantity variables (number of time slots, number of complete diaries) between modes, before interviewer correction, using the differences between modes for each person
- Estimate overall mode effects for quantity variables with a matched-pairs approach between participants randomly assigned to PAPI or CAWI as the first mode.
- Comparison of information precision (duration differences, weighted comparison with population, use of defaults), before interviewer correction, accounting for order effect and non-response
- Establish predictors in both modes that predict high data quality in the other mode (relative benefit of interviewer correction)

Additional considerations:

The experimental design relies on each person forming their own control across modes. This can lead to increased concerns with respect to non-response and/or dropout. Households least comfortable with filling their activities online (as identified by proxy variables of internet competence and preferences) are thus less likely to respond via the CAWI mode if given the choice. Analyses are planned to identify subgroups likely to be non-responders within each mode.

The design of France's TUS app offers a unique opportunity to investigate the consequences of different navigational paths selected by respondents; they may click through a series of hierarchical categorical buttons represented with pictograms, they may type in an activity which searches a database, or they may elect to enter their activity as free text. Using different navigational mechanisms may impact the data's accuracy but also lead to distinct differences between modes. A per-category analysis of the impact of respondent path selection is planned to identify impactful aspects of this smart feature.

France HBS

Research question(s):

- How does mode of administration impact measures of household consumption?
- Do participants who choose the paper mode of administration differ from those who choose the smart survey mode?
- To what degree do non-response and measurement mode effects interact on quantity and quality of data?
- Is the amount of information provided in the app diary equal to that of the paper diary?

- Are the features of the diary app easy for the respondent to use while completing the survey?

How to answer these:

- Estimate overall mode effects for quantity variables (number of expenses, number of receipts in the diary, response rates in the questionnaire) between modes
- Use of paradata of the app (for example: unsuccessful attempts to add a product in the list are recorded)

Additional considerations:

For the different collection tools selected by the household, dropout rates will be measured at different stages of the survey.

A specific module in the second questionnaire allows the respondent to evaluate the different features of the app (from easy to use to difficult to use)

Belgium

Research question(s):

- Can data from PAPI follow-up responders be integrated in a mixed-mode manner?
- How does mode of administration impact measures of time use?

How to answer these:

- Analysis of integral quality/quantity variables with sequential mixed-mode design, with instrumental variable approach (Vannieuwenhuyze, Loosveldt and Molenberghs (2010)
- Compare with multi-source analysis, in which quality/quantity variables are estimated within each source.

Additional considerations:

In Belgium, unlike in France, the comparison group opting into the paper survey is unlikely to be very large, as it represents a subset of half of the non-responders. Because of this, both non-response and measurement differences will be embedded in the mode effect within this study, and they may be difficult to parse out. This, in combination with the larger measurement differences, may reduce the efficacy of a mixed-mode data integration approach. The approach would then be to get separate estimates from the app and PAPI modes and combine them. It may be feasible to use the information on pure mode effect/measurement mode effect arising from the French study to inform these estimates.

The TUS app deployed in this field test involved a hard cutoff for interaction at the end of a given number of days days for data privacy considerations. In comparison, it is often impossible to determine when a person completed the equivalent paper diary.

Belgium will link diary responses to administrative data at 1) the household level, including household type (in 7 categories), region (in 3 categories), net income (in deciles), urbanicity (semi-rural to urban), and home-ownership; 2) the individual level, including gender, age, position in household (in 7 categories), country of birth (in 3 categories), social-economic position (in five categories), education (grouped according to the ISCED classification), and income (in deciles.) These data will be used as covariates in assessing differential non-response between the two modes.

Italy

Research question(s):

- Does prepopulating the diary using geolocations impact the content of the TUS diary?
- Does prepopulating the diary using geolocations impact measures of quality of the TUS diary?

How to answer these:

- Analysis of the experiment (data quality/data quantity), using personal covariates to adjust for residual representation differences
- Investigation of drop-out between the two groups

Additional considerations:

In Italy, because both groups will be recruited face-to-face by an interviewer and both groups are using the same app with a more-smart and less-smart set of smart features, the expectation is that representational differences will be largely non-existent.

4.5 Conclusion

This chapter describes the current efforts of task 2.4 to provide a way forward for integrating smart and non-smart surveys. The question of how we compare and integrate data across survey modes that may differ substantially following the introduction of smart features is not limited to the moment in which countries introduce Smart Surveys. Instead, it is likely to remain relevant for the foreseeable future as researchers mitigate concerns of representativity.

The experiments conducted in France, Belgium, and Italy as part of their Household Budget Survey and Time Use Survey field tests will provide data that will permit estimation of measurement and representational differences. The analysis of this data should result in functional guidelines for assessing the feasibility of simple integration strategies. Additionally, the investigation of specific measurement differences in the form of improvements in data quality will offer insight into the benefits of smart feature sets to correct traditional survey techniques at the point of integration. By providing guidelines and strategies for tackling these issues, the findings of task 2.4 will contribute to researchers' capacity to utilize new technologies without sacrificing statistical rigor.

References:

Bauman, A., Bittman, M., & Gershuny, J. (2019). A short history of time use research; implications for public health. BMC Public Health, 19(Suppl 2), 607. https://doi.org/10.1186/s12889-019-6760-y

Bucher Della Torre, S., Carrard, I., Farina, E., Danuser, B., & Kruseman, M. (2017). Development and evaluation of e-CA, an electronic mobile-based food record. Nutrients, 9(1), 76. https://doi.org/10.3390/nu9010076

Chatzitheochari, S., Fisher, K., Gilbert, E., Calderwood, L., Huskinson, T., Cleary, A., & Gershuny, J. (2018). Using new technologies for time diary data collection: Instrument design and data quality findings from a mixed-mode pilot survey. Social Indicators Research, 137(1), 379–390. https://doi.org/10.1007/s11205-017-1569-5

Conrad, F. G., Couper, M. P., Tourangeau, R., & Galesic, M. (2005). Interactive Feedback Can Improve the Quality of Responses in Web Surveys.

Frąckowiak, M., Rogowski, Ł., & Sommer, V. (2022). Hopes and challenges of creating and using a smartphone application. Working on and working with a digital mobile tool in qualitative sociospatial research. Qualitative Research: QR, 146879412210989. https://doi.org/10.1177/14687941221098923

Gillis, D., Lopez, A. J., & Gautama, S. (2023). An evaluation of smartphone tracking for travel behavior studies. ISPRS International Journal of Geo-Information, 12(8), 335. https://doi.org/10.3390/ijgi12080335

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., & Crane, M. (2015). A Web-Based Diary and Companion Smartphone app for Travel/Activity Surveys. Transportation Research Procedia, 11, 297–310. https://doi.org/10.1016/j.trpro.2015.12.026

Harms, T., Gershuny, J., Doherty, A., Thomas, E., Milton, K., & Foster, C. (2019). A validation study of the Eurostat harmonised European time use study (HETUS) diary using wearable technology. BMC Public Health, 19(Suppl 2), 455. https://doi.org/10.1186/s12889-019-6761-x

Jäckle, A., Gaia, A., & Lessof, C. (2019). A review of new technologies and data sources for measuring household finances: Implications for total survey error. Working Papers for a New Society.

Kelly, P., Thomas, E., Doherty, A., Harms, T., Burke, Ó., Gershuny, J., & Foster, C. (2015). Developing a method to test the validity of 24 hour time use diaries using wearable cameras: A feasibility pilot. PloS One, 10(12), e0142198. https://doi.org/10.1371/journal.pone.0142198

Keusch, F., Bach, R., & Cernat, A. (2022). Reactivity in measuring sensitive online behavior. Internet Research, 33(3), 1031–1052. https://doi.org/10.1108/intr-01-2021-0053

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2023). Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey. Sociological Methods & Research, 52(2), 841–878. https://doi.org/10.1177/0049124120914924

Klausch, T., Hox, J., & Schouten, B. (2015). Selection error in single-and mixed mode surveys of the Dutch general population. Journal of the Royal Statistical Society. Series A, 178(4).

Klingwort, J., & Schnell, R. (2020). Critical Limitations of Digital Epidemiology: Why COVID-19 Apps Are Useless. Survey Research Methods, 14(2), 95–101. https://doi.org/10.18148/srm/2020.v14i2.7726

Lev-On, A., & Lowenstein-Barkai, H. (2019). Viewing diaries in an age of new media: An exploratory analysis of mobile phone app diaries versus paper diaries. Methodological Innovations, 12(1), 205979911984444. https://doi.org/10.1177/2059799119844442

Lunardelli, I., van den Heuvel, J., Schouten, B., Nuccitelli, A., D'Amen, B., Lorè, B., Perez, M., & Zgonec, M. (2024). *Deliverable 1.2: Smart Baseline stage report ANNEX: Smart perceptions survey* (Deliverable No. 1.2). CBS, The Netherlands; ISTAT, Italy; SURS, Slovenia. URL

Premkumar, P. S., Ganesan, S. K., Pandiyan, B., Krishnamoorthy, D. K., & Kang, G. (2023). Smartphone diary application in household surveys: Integration of high frequency temporal data in large-scale data collection. Field Methods, 1525822X231195525. https://doi.org/10.1177/1525822x231195525

Riegler, M. P. R. (2015). Integrating the Web Mode in. cros-legacy.ec.europa.eu.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. Social Science Research, 42(6), 1555–1570. https://doi.org/10.1016/j.ssresearch.2013.07.005

Schwarz, N. (2012). Why researchers should think “real-time”: A cognitive rationale. dornsife.usc.edu.

Sullivan, O., Gershuny, J., Sevilla, A., Walthery, P., & Vega-Rapun, M. (2020). Time use diary design for our times - an overview, presenting a Click-and-Drag Diary Instrument (CaDDI) for online application. Journal of Time Use Research, 1–17. https://doi.org/10.32797/jtur-2020-1

Tourangeau, R. (2017). Mixing modes. In Total Survey Error in Practice (pp. 115–132). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119041702.ch6

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. Public Opinion Quarterly, 74(5), 1027–1045. https://doi.org/10.1093/poq/nfq059

Conclusion

This deliverable presents the work that is planned in work package 2 of the Smart Survey Implementation project, and intermediate findings. The small and large field tests that are conducted in 2024 will be important in answering the central questions related to the four main tasks in this project

1. The successful **recruitment** of participants for smart surveys.
2. Using **machine learning** to improve Human-Computer Interaction in smart surveys.
3. Respondent involvement and **human-computer interaction** in smart surveys.
4. Integrating smart surveys with traditional survey methods by estimating the mode e**ffect.**

In this deliverable, Chapter 1 illustrated how a series of field experiments will be used to test various designs that should enable the successful recruitment of respondents into smart surveys. Chapter 2 showed how machine learning will be used in the microservices developed in work package 3 to process smart data, and feed information back to the respondent. Chapter 3 presented the protocol that will be used to test the end-to-end procedures developed in the project for conducting smart surveys. Chapter 4 explained how several parallel runs and experiments will be used to estimate the size of the mode effect when a traditional surveys transitions to a smart survey.

It is clear that the tests conducted in 2024 will be very important to inform successful strategies for smart surveys. The results of these tests will be reported in the final work package 2 deliverable that is foreseen for April 2025.

Appendix A: detailed information on auxiliary geodata and GPS sensors used in the geolocation microservice for Time Use Surveys.

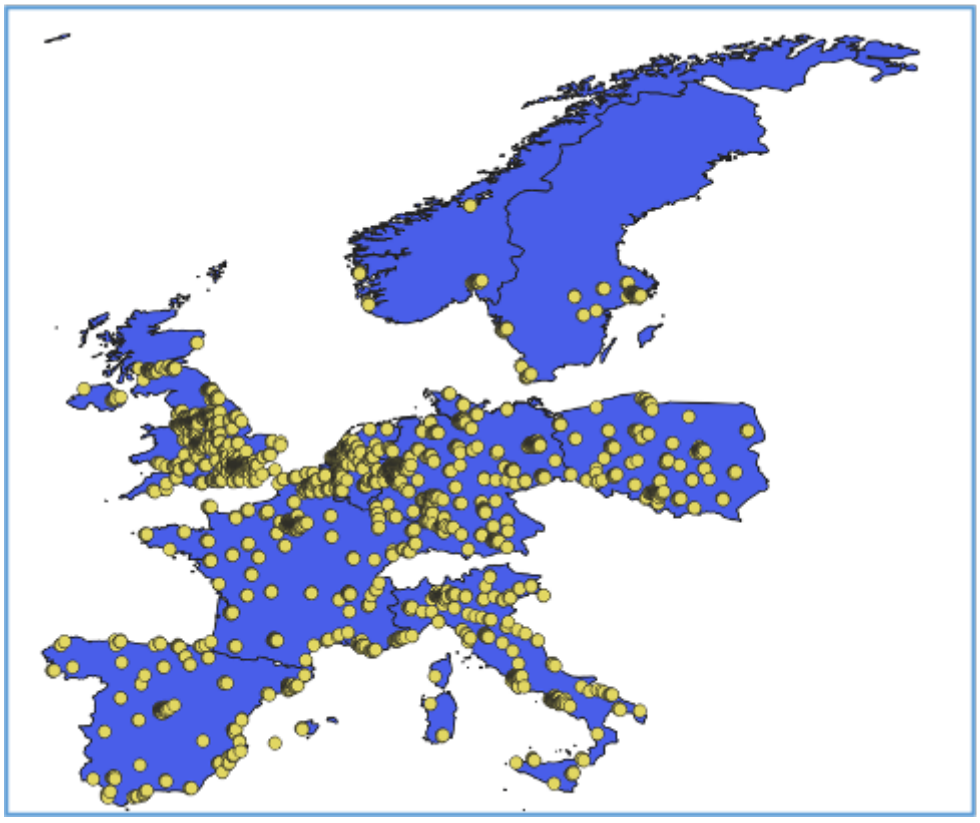
Comparison between Google Places and Open Street Maps (OSM)

The "Comparison between Google Places and Open Street Maps (OSM)" analyses the quality and coverage of Points of Interest (POI) between these two platforms.

Using a methodology that involves randomly selecting points within urban areas of 11 countries, the study compares POI data collected within a 50-meter radius around each point.

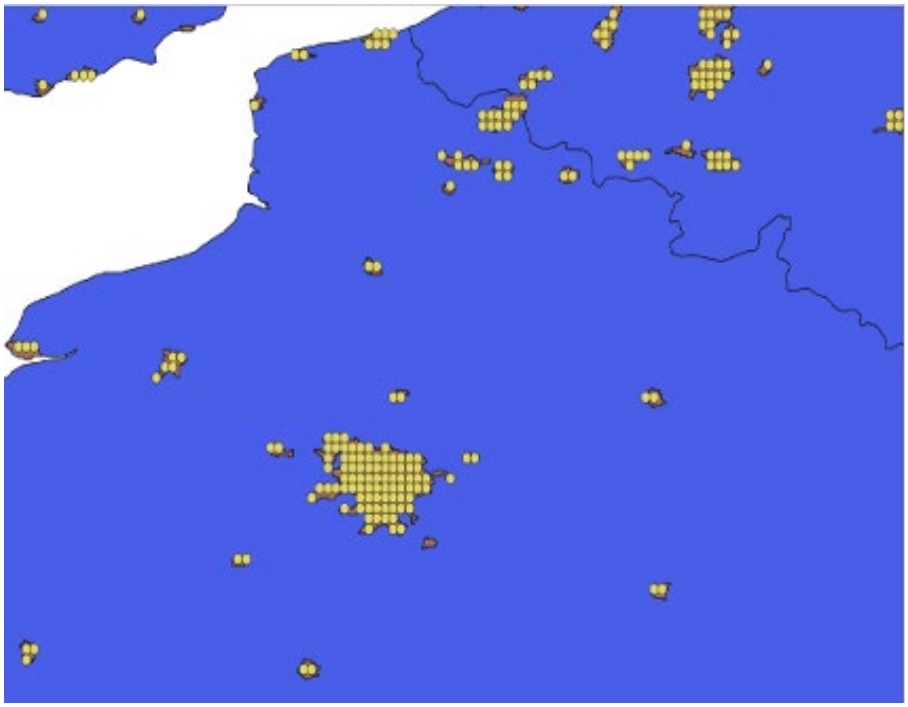
The results show that Google Places identified 7,831 POIs compared to 1,401 by OSM. Although OSM's coverage is generally lower, it includes unique POIs not found in Google Places. However, only 221 out of the 1,401 OSM POIs matched with Google Places POIs, indicating significant differences in POI representation between the platforms. The study highlights the need to combine or compare different POI sources for a comprehensive understanding of POI distribution in urban areas.

- Methodology: random arrows on 11 countries
- Quality and coverage comparison of Points of Interest (POI) between Google Places (GP) and Open Street Maps (OSM).
- The adopted methodology involved the random selection of a appropriate number of points (similar to throw *arrows*) across urban areas with populations exceeding 250,000 inhabitants, in the 11 countries promoting the initial ESSNET Smart Surveys project.
- Urban areas' choice has been adopted to exclude uninhabited areas like seas, lakes, and mountains.



Methodology: radius choice and grid

- POI detection was limited to a 50-meter radius around each selected point, due to Google Places' restrictions on the maximum number of retrievable POIs, capped at 20.
- The points were distributed using a 10km grid overlay on a GIS map.
- The grid adoption guarantees to allocate several points to the larger urban areas, as it happens in the following picture regarding Paris:



First results

- In total, **2223** arrows were analyzed. POIs were collected within a 50-meter radius. At the end of the analysis, **Google Places** yielded a total of **7831** POIs, while **Open Street Map** identified **1401**.
- In OSM, only POIs with specific tags were considered, specifically: *amenity, office, shop, healthcare, land use, craft, tourism, leisure*.
- However, it was noted that despite the lower number of POIs in OSM, some of these (such as rest benches or public fountains) might not be directly relevant for comparison with Google Places POIs.
- POI coverage varies significantly across countries.

Table A1: Google Places (GP) and OpenStreetMaps information.

ISO2	GP_pois	OSM_pois	coverageOSM	arrows	GP_hits	OSM_hits	countmerged
DEU	1689	376	22,3%	458	3,69	0,82	56
FRA	1119	311	27,8%	297	3,77	1,05	45
GBR	1814	224	12,3%	576	3,15	0,39	34

ITA	916	136	14,8%	233	3,93	0,58	27
ESP	588	110	18,7%	187	3,14	0,59	17
POL	654	96	14,7%	176	3,72	0,55	16
NLD	532	60	11,3%	129	4,12	0,47	18
SWE	188	39	20,7%	61	3,08	0,64	3
BEL	235	28	11,9%	76	3,09	0,37	3
LUX	24	16	66,7%	3	8,00	5,33	1
NOR	72	5	6,9%	27	2,67	0,19	1
total	**7831**	**1401**		**2223**			**221**

- OSM coverage is consistently lower than that of GP (coverageOSM), and
- often the number of POIs found in OSM is even less than the number of points analyzed per country (arrows).
- The average number of POIs detected per point is always greater than 1 in GP (GP_hits),
- but less than 1 in OSM (OSM_hits), except for Luxembourg, where OSM coverage is notably high.

Remarks

- OSM POIs are not a subset of those found in Google Places, not even partially.
    - Through a record linkage process, only 221 out of 1401 POIs could be matched between the two sources, increasing to 296 when relaxing linkage criteria.
    - This highlights significant differences in POI representation between the two platforms, despite efforts to ensure comparability between OSM tags and GP labels.
    - The matching seems to be different regarding different kinds of POIs (e.g. shops better than offices).
- The results of this analysis, although exploratory and conducted at a macro level, underscore significant quantitative and qualitative differences between Google Places and Open Street Map.
- Also, if we were playing with GP we got in trouble: our Maps API has been restricted for 24 hours. Let's think about working (and how) on this platform.
- This suggests that combining or comparing different POI sources may be necessary for a comprehensive understanding of POI distribution in urban areas.

Analysis of the quality of GPS sensor across smartphone types and countries

The following tables show the ranking of the five best-selling smartphone models in some European countries (Italy, Germany, Belgium, Netherlands, Spain and Slovenia), with an indication of the type of GPS sensor and a measurement of accuracy. Furthermore, the following figures show the distributions of smartphone brands in the same countries.
These distributions highlight remarkable differences across countries, which may impact the performance of GPS data processing.

Table A2: smartphone model, GNSS system, sensors, maximum accuracy in meters, current price, ranking position by country (year specified in parentheses)

Model	GNSS	Sensors	Max orbit precision (m)	Price (€)	Popularity rank in various countries				
					IT	DE	BE	NL	SI
Samsung Galaxy A54	GPS, GLONASS, GALILEO, BDS	accelerometer, gyro, compass, barometer	A+ (about 2m) (1) Samsung Galaxy A53 5G	299				**1**	
Iphone 13	GPS, GLONASS, GALILEO, BDS, QZSS	accelerometer, gyro, proximity, compass, barometer	About 3.44 meters (2)	669	**1**	**1**	**3**	**2**	**4**
Iphone 15	GPS, GLONASS, GALILEO, BDS, QZSS	accelerometer, gyro, proximity, compass, barometer	About 3.44 meters (2) Iphone 13	899				**3**	
Iphone 14	GPS, GLONASS, GALILEO, BDS, QZSS	accelerometer, gyro, proximity, compass, barometer	About 3.44 meters (2) Iphone 13	779	**2**	**3**	**2**	**4**	
Samsung galaxy a14	GPS, GALILEO, GLONASS, BDS, QZSS	accelerometer, proximity, compass	About 3.44 meters (2) Iphone 13	159			**4**	**5**	
Iphone 14 Pro/Pro max	GPS (L1+L5), GLONASS, GALILEO, BDS, QZSS	accelerometer, gyro, proximity, compass,barometer	About 3.44 meters (2) Iphone 13	889	**3**	**4**	**1**		
Samsung Galaxy S23 Ultra	GPS, GLONASS, BDS, GALILEO	accelerometer, gyro, proximity, compass, barometer	About 2.95 meters (2) Samsung Galaxy S21	999			**5**		

iPhone 11	GPS, GLONASS, GALILEO, QZSS	accelerometer, gyro, proximity, compass, barometer	About 3.29 meters (2)	569	**2**	**2**
Iphone 12	GPS, GLONASS, GALILEO, QZSS	accelerometer, gyro, proximity, compass, barometer	About 2.73 meters (2)	599	**5**	**5**
Samsung Galaxy A12	GPS, GLONASS, GALILEO, BDS	accelerometer	A- (4-5m) (1)	149		
Xiaomi redmi note 10 Pro	GPS, GLONASS, GALILEO, BDS	accelerometer, gyro, compass	B (8-9m) (1) Xiaomi redmi note 9 Pro	199		
Xiaomi redmi 9a	GPS, GLONASS, BDS	Accelerometer, proximity	B (8-9m) (1) Xiaomi redmi 8a	79		
Xiaomi redmi Note 10S	GPS, GLONASS, GALILEO, BDS	accelerometer, gyro, compass	B (8-9m) (1) Xiaomi redmi note 9 Pro	199		
Oppo a53S	GPS, GLONASS, GALILEO, BDS, QZSS	accelerometer, gyro, proximity, compass	B+ (6-7m) (1)	170		
Samsung Galaxy A52 5G	GPS, GLONASS, GALILEO, BDS	accelerometer, gyro, compass	A- (4-5m) (1)	299		**1**
Samsung Galaxy A53 5G	GPS, GLONASS, GALILEO, BDS	accelerometer, gyro, compass, barometer	A+ (about 2 m) (1)	299		**3**
Galaxy a52s 5G	GPS, GLONASS, GALILEO, BDS, QZSS	accelerometer, gyro, compass	B- (10-12m) (1)	269	**4**	
Galaxy A22 5G	GPS, GLONASS, GALILEO, BDS	accelerometer, gyro, proximity, compass	B+ (6-7m) (2)	268	**5**	

Samsung
Galaxy A23

Notes: The data relating to the closest model are shown in red, due to a lack of information on the selected model. Most popular brands are shown for IT: 2022-2023, DE: 2021-2023, BE: 2022-2023, NL: 2023, SI: 2022

Sources of information:

1. http://www.spillby.com/gpstest/index.php?p=5
2. https://www.fs.usda.gov/database/gps/mtdcrept/accuracy/documents/2023%20GNSS%20Test%20Data%20Report_1_13_2023.pdf
   (nota metodologica: https://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3)

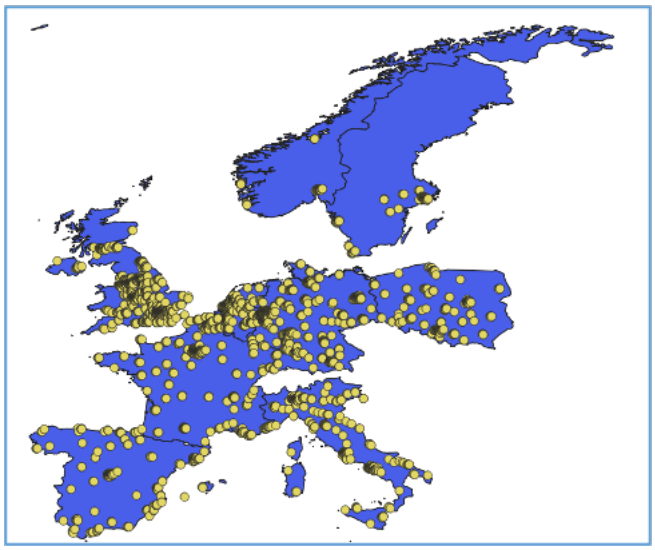


Figure: 2.1 Smartphone brands by country

[1] https://urlsand.esvalabs.com/?u=https%3A%2F%2Fcros.ec.europa.eu%2Fbook-page%2Freport-methodology-21&e=2637b01d&h=0b87d861&f=y&p=y

Appendix B: country documentation for Belgium

Methodological Test of TUS

Date: March/April 2024
Net sample: 6,000 dwellings, selected randomly from the population register, proportionally across all three Belgian regions. From each sampled dwelling, a single reference person was chosen at random.
Modes: app-based TUS/CAWI, PAPI
Incentive structure: 15 euros conditional upon completion of the survey

Protocol

Invitation by postal mail sent to all sampled households requesting participation in an app-based Time Use Study (MOTUS). Non-responders are followed up at 2 and 4 weeks. MOTUS participants complete a preliminary profile questionnaire, followed by seven days of time use registration, followed by a closing questionnaire on leisure and sports activities in the previous week. Participants completing the PAPI diary completed seven days of time use registration, followed by the profile questionnaire and closing questionnaire on health, leisure and sports activities in the previous week, and questions related to the filling-in of the diary.

Detailed protocol information:

Aim: Investigate the impact of invitation design, follow-up prompting, and subsequent mode measurement differences.

Experimental conditions

Each sampled household is contacted by post delivered to their registered address by the Belgian National Statistical Institute (StatBel). The sample is randomly allocated to one level in two different experimental conditions in a 2x2 between groups design. The first experiment (*Design*) concerns differences in tone, layout, and engagement approach of the invitation letters. Sampled households in the traditional letters condition received an invitation written in a more formal style, focusing on the official nature of the study and the importance of cooperation. Households allocated to the user-friendly condition received a letter with a more casual, conversational tone, emphasizing the personal benefit and impact of participation. Figure A.1 demonstrates the extent of the layout differences between the two letters.

Figure A.1

The second experiment (*PAPI follow-up*) concerns the offer of alternative mode at the first follow-up moment. All non-responding households were contacted with a letter reminding the household that they had been invited to participate in the survey that also included their MOTUS username and password. Non-responding households allocated to the PAPI follow-up condition were also sent the full PAPI diary along with a self-addressed, prepaid envelope. The follow-up letters in this condition contained an additional section at the end explaining how to complete and return the paper diary should the household prefer this version. These reminder letters differed in content according to the Design condition allocation of the household, but both the traditional and user-friendly conditions received identically worded explanations concerning the PAPI option.

MOTUS

All invitation letters contained information on downloading and logging into MOTUS. MOTUS was made available to participants as a web application, an iOS app published in the Apple store, and as an Android app published in the Google Play store. After downloading the app or visiting the site, respondents activate their account using the code provided in the invitation letter and linking it to an email address. Within MOTUS, users had access to instructional materials and general information on the study.

Respondents to the MOTUS mode received email reminders to motivate them to start or complete the next required activity.

Additional variables

Diary responses will be linked to administrative data at 1) the household level, including household type, region, net income, urbanicity, and home-ownership; 2) the individual level, including gender, age, position in household, country of birth, socioeconomic position, education, and income.

Research questions

1. Recruitment
    - How does the overall design of the invitation letter affect the response rate?
    - How does the inclusion of a follow-up PAPI mode affect the response rate in a smart survey?
2. Mode effects
    - Can data from PAPI follow-up responders be integrated in a mixed-mode manner?
    - How does the mode of administration impact measures of time use?

Materials for fieldwork Belgium

STATBEL
België in cijfers

Datum

Algemene Directie Statistiek - Statistics Belgium
Koning Albert II-laan 16 – 1000 Brussel

{{voornaam_contact}} {{naam_contact}}
{{TX_ADRESS}}
{{CD_ZIP_RES}} {{TX_city}}

Uw berichten	Uw kenmerk	Ons kenmerk	Bijlagen
	{{nr_ID}}		

Betreft: Tijdsbestedingsonderzoek 2024

Geachte {{mevrouw/heer}} {{naam_contact}},

We nodigden u een tijdje geleden uit om deel te nemen aan een studie over tijdsbesteding. Als u de vragenlijsten en het dagboekje al invulde, bedanken wij u daarvoor. Kwam u er nog niet toe, dan kan u dit nog steeds doen.

De kwaliteit van ons onderzoek steunt op uw medewerking. Een vergoeding van **15 euro** is voorzien voor de tijd en de energie die u hieraan wilt besteden.

De studie verloopt online en bestaat uit drie onderdelen.

1. U vult een vragenlijst in over uw achtergrond, uw opleiding, uw werk, uw tijdsbesteding en uw mening daarover.
2. U houdt uw dagelijkse activiteiten bij in een online dagboekje gedurende 1 week.
3. U beantwoordt nog enkele vragen over de week waarin u uw dagboekje bijhield en uw deelname aan vrijetijds- en sportactiviteiten.

U kunt de vragenlijsten invullen en het dagboekje bijhouden via een webapplicatie op uw computer, laptop of tablet of via een app op uw smartphone of tablet.

Hiervoor werkt Statbel samen met de onderzoeksgroep TOR van de Vrije Universiteit Brussel. Zij ontwierpen het software platform MOTUS dat voor deze studie wordt gebruikt.

Om mee te doen, downloadt u de MOTUS-app van hbits of scant u de QR-code hieronder of surft u naar **www.motusresearch.io/nl** en vul uw gebruikersnaam en wachtwoord in.

Uw gebruikersnaam: **{{respondent_username}}**
Uw wachtwoord: **{{respondent_password}}**

Contactpersoon dienst Tijdsbestedingsonderzoek
Algemene Directie Statistiek - Statistics Belgium | TUS@economie.fgov.be | + 32 800 120 33

Ondernemingsnummer: 0314.595.348

economie
FOD Economie, K.M.O., Middenstand en Energie

+32 800 120 33 | Statbel.NL | @Statbel_nl | /statbel | statbel.fgov.be | .be

STATBEL
België in cijfers

{{voornaam_contact}} {{naam_contact}}
{{TX_ADRESS}}
{{CD_ZIP_RES}} {{TX_city}}

Geachte {{voornaam_contact}} {{naam_contact}},

We nodigden u een tijdje geleden uit om deel te nemen aan een studie over tijdsbesteding. **Als u de vragenlijsten en het dagboekje al invulde, bedanken wij u daarvoor.** Kwam u er nog niet toe, dan kan u dit nog steeds doen.

We belonen deelnemers!
De studie bestaat uit het beantwoorden van twee vragenlijsten en het bijhouden van uw dagelijkse activiteiten in een dagboekje gedurende 1 week. Voor de deelname aan het volledige onderzoek, ontvangt u een belastingvrije vergoeding van 15 euro.

Meedoen is makkelijk & vertrouwelijk

U kunt de vragenlijsten invullen en het dagboekje bijhouden via een app of een webtoepassing:

- download de MOTUS-applicatie van hbits op uw smartphone van de Apple App Store of van de Google Play Store, afhankelijk van uw besturingssysteem; of
- surf naar **www.motusresearch.io/nl** of scan de QR-code hiernaast met uw smartphone of tablet

Log in met volgende gebruikersnaam en wachtwoord:
Uw gebruikersnaam: **{{respondent_username}}**
Uw wachtwoord: **{{respondent_password}}**

Verkiest u op papier deel te nemen? Vul het dagboek en de vragenlijst in en stuur deze gratis terug in de voorgedrukte, geadresseerde en gefrankeerde envelop.

We respecteren uw privacy en we verwerken uw gegevens vertrouwelijk en anoniem. Lees onze privacyverklaring op https://statbel.fgov.be/nl/privacy. Het is ons alleen om gemiddelden en statistieken te doen.

Heeft u nog vragen?
Check onze FAQ op https://statbel.fgov.be/nl/enquete/onderzoek-naar-tijdsbudget-tus
Mail naar TUS@economie.fgov.be
Bel naar het gratis nummer 0800 120 33 van de FOD Economie.

Hartelijk dank voor uw medewerking!
Met vriendelijke groeten,

Philippe Mauroy
Directeur-generaal a.i. Statbel

Algemene Directie Statistiek - Statistics Belgium
Koning Albert II-laan 16 – 1000 Brussel

Datum

{{voornaam_contact}} {{naam_contact}}
{{TX_ADRESS}}
{{CD_ZIP_RES}} {{TX_city}}

Uw berichten	Uw kenmerk	Ons kenmerk	Bijlagen
	{{nr_ID}}		

Betreft: Tijdsbestedingsonderzoek 2024

Geachte {{mevrouw/heer}} {{naam_contact}},

U bent samen met zo'n achtduizend andere personen geselecteerd voor een onderzoek over de tijdsbesteding van de Belgen. Ik nodig u in naam van Statbel, het Belgische statistiekbureau, uit om deel te nemen aan deze studie.

De kwaliteit van ons onderzoek steunt op uw medewerking. De deelname aan dit onderzoek is vrijblijvend. Een vergoeding van **15 euro** is voorzien voor de tijd en de energie die u hieraan wilt besteden.

De vragen gaan over uw dagelijks leven, met thema's zoals betaald en huishoudelijk werk, onderwijs, vrije tijd, mediagebruik, sociale contacten en verplaatsingen. Uw antwoorden zullen worden gebruikt om het sociaal beleid van België en Europa mee vorm te geven.

De studie verloopt online en bestaat uit drie onderdelen.

1. U vult een vragenlijst in over uw achtergrond, uw opleiding, uw werk, uw tijdsbesteding en uw mening daarover.
2. U houdt gedurende 1 week uw dagelijkse activiteiten bij in een online dagboekje.
3. U beantwoordt nog enkele vragen over de week waarin u uw dagboekje bijhield en uw deelname aan vrijetijds- en sportactiviteiten.

U kunt de vragenlijsten invullen en het dagboekje bijhouden via een webapplicatie op uw computer of tablet of via een app op uw smartphone.

Hiervoor werkt Statbel samen met de onderzoeksgroep TOR van de Vrije Universiteit Brussel. Zij ontwierpen het software platform MOTUS dat voor deze studie wordt gebruikt.

Om mee te doen, downloadt u de MOTUS-app van hbits of scant u de QR-code hieronder of surft u naar **www.motusresearch.io/nl** en vul uw gebruikersnaam en wachtwoord in.

Uw gebruikersnaam: **{{respondent_username}}**
Uw wachtwoord: **{{respondent_password}}**

Contactpersoon dienst Tijdsbestedingsonderzoek
Algemene Directie Statistiek - Statistics Belgium | TUS@economie.fgov.be | + 32 800 120 33

Ondernemingsnummer: 0314.595.348

STATBEL
België in cijfers

{{voornaam_contact}} {{naam_contact}}
{{TX_ADRESS}}
{{CD_ZIP_RES}} {{TX_city}}

Helpt u ons een beeld te krijgen hoe Belgen hun tijd besteden?

Geachte {{voornaam_contact}} {{naam_contact}},

Hoeveel tijd besteedt u aan verplichtingen zoals betaald werk, opleiding en zorgtaken? Hoe vaak en hoe lang bent u bezig met persoonlijke verzorging? Hoeveel tijd blijft er over voor ontspanning? ... Statbel, het Belgische statistiekbureau, doet onderzoek naar het tijdsgebruik van de Belg en werkt hiervoor samen met hbits en de onderzoeksgroep TOR van de Vrije Universiteit Brussel.

Waarom deelnemen?

Meedoen aan het tijdsbestedingsonderzoek is een unieke kans om de dagelijkse realiteit te laten zien. U zorgt ervoor dat uw ervaringen worden meegenomen in het beeld dat we krijgen van onze samenleving en op die manier kan u helpen om diensten te verbeteren en richting te geven aan het beleid.

We belonen deelnemers!

De studie bestaat uit het beantwoorden van twee vragenlijsten en het bijhouden van uw dagelijkse activiteiten in een dagboek gedurende één week. Voor de deelname aan het volledige onderzoek, ontvangt u een belastingvrije vergoeding van 15 euro.

Hoe neemt u deel?

U kunt de vragenlijsten invullen en het dagboek bijhouden via een app of een webtoepassing:

- download de MOTUS-applicatie van hbits op uw smartphone van de Apple App Store of van de Google Play Store, afhankelijk van uw besturingssysteem; of
- surf naar **www.motusresearch.io/nl** of scan de QR-code hiernaast met uw smartphone of tablet

Log in met volgende gebruikersnaam en wachtwoord:

Uw gebruikersnaam: **{{respondent_username}}**
Uw wachtwoord: **{{respondent_password}}**

We respecteren uw privacy en we verwerken uw gegevens vertrouwelijk en anoniem. Lees onze privacyverklaring op https://statbel.fgov.be/nl/privacy. Het is ons alleen om gemiddelden en statistieken te doen.

Heeft u nog vragen?

Check onze FAQ op https://statbel.fgov.be/nl/enquete/onderzoek-naar-tijdsbudget-tus

Mail naar TUS@economie.fgov.be

Bel naar het gratis nummer 0800 120 33 van de FOD Economie.

Hartelijk dank voor uw medewerking!

Met vriendelijke groeten,

Philippe Mauroy, Directeur-generaal a.i. Statbel

Appendix C: country documentation for France

Methodological test of TUS by Insee

Date: Q2 2023 (9 May to 1 July 2023)
Sample size: 2,100 dwellings
Respondents: 627 dwellings / 1070 respondents (597 with both a PAPI and CAWI diary, 198 with only a PAPI diary, 75 with only a CAWI diary, 200 with none).

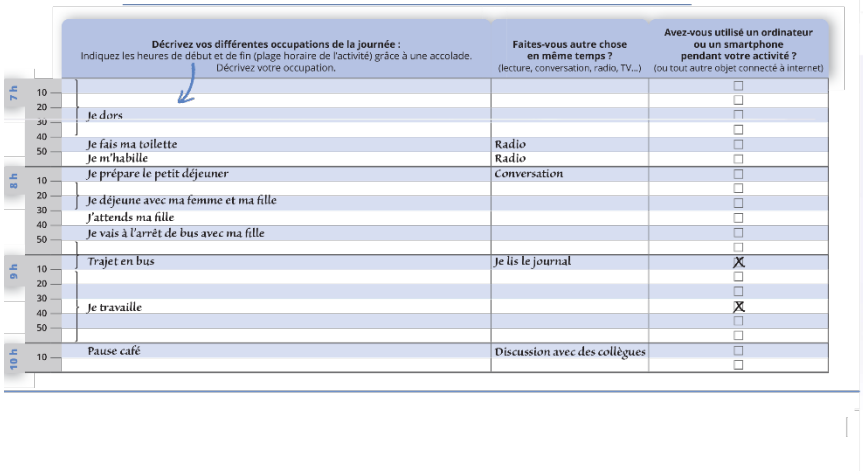
Protocol: 1st visit of interviewer / filling of two types of diaries / 2nd visit of interviewer
Sample with varied profiles, randomly divided into 2 sub-samples (for the 2 orders of filling in the diaries, numeric first or paper first).
Restriction: Households must have access to the internet to be able to respond and be sufficiently comfortable with the internet.

Aim:
- Measuring the impact of the digital versus paper diary
- Acceptability of the survey and the protocols, for the respondent but also for the interviewer

Detailed protocol:
The aim is to compare the descriptions of the comparable time periods in both the electronic and paper diaries, without the recording in one mode having too much influence on the recording in the other, due to the learning or fatigue effect. The protocol devised to approach this ideal situation is to have the same respondent repeat the survey over two time periods that are as close as possible: the same day of the week at a week's interval seems to be the best compromise. This involves carrying out a test-retest consisting of, for each respondent, describing a day (for example, Tuesday) within the assigned mode, and then describing the same day the following week (still Tuesday), within the other mode. To control for the sequence effect, part of the sample will complete this entry in PAPI -> CAWI order, the other in reverse order (CAWI -> PAPI). To collect the paper diaries, a second face-to-face visit is necessary. During this visit respondents are asked about their difficulties in filling their activities in both modes, and about their preferred mode. Figure B.1 provides an illustrative example of differences in format between the CAWI and PAPI modes.

Figure B.1

a) Example filled PAPI diary



a) Example filled CAWI diary

As the two days may differ objectively from each other, care should be taken by the interviewer to collect general information on the type of day in terms of time use (ordinary, more complex, etc.) when describing the second day.

Several individuals are interviewed in the household: individual, possible spouse, randomly selected child aged 11 to 24.

Research Questions in TUS

Mode Effects

- Is the quantity of information provided in the digital diary equal to that provided by the paper diary (number of time slots, number of complete diaries...)?
- How does the precision of this information differ between modes? What is the difference between the average duration of the various activities (work time, leisure time, sleep time, domestic work time...) between the paper and digital diaries?
    - Does the specific CAWI path used for inputting activities improve respondent's precision?
- To what degree do non-response and measurement mode effects interact on the quantity and precision of data?

UX/UI impact

- How can one maximize the accuracy of information provided by the CAWI diary? (e.g., through improvements of the app's design, of activities nomenclature labels...)
- How does the navigational path chosen by respondents (categorical buttons, search field, open field) impact the ability to select the correct activities, and in what way does it interact with the type of chosen activity?

- Acceptability, motivation, qualityWhat mode do respondents prefer to use, and how does this impact their completion of the diary?
- What respondent characteristics can predict respondent's mode preference?
- Analysis of questions asked to the interviewer about his or her work in checking the diaries (web and paper) with the household

Methodological test of HBS by Insee

Date: Q2 2024
Sample: 2,400 households (Mandatory for respondents), with an expected response rate of 55% Protocol:
- subsample 1 (500 dwellings): 1st visit of interviewer / filling of paper diary / 2nd visit of interviewer
- subsample 2 (1,900 dwellings): 1st visit of interviewer / choice of paper or app-based diary / filling of a web questionnaire / 2nd visit of interviewer (shorter if web questionnaire filled in)
Restriction: Elderly people not included.

Detailed protocol:
The aim is to compare the historical protocol of HBS survey in France to the expected protocol of HBS 2026, which will allow the respondent to choose between completing an app-based diary or a paper diary.
The app-based diary is HBS app, developed during past years by CBS.
To test the mode effect of the diary in real conditions, the respondent has to go through the whole survey, which consists of an initial face-to-face interviewer visit to the household, followed by the completion of a seven-day expenditure diary, followed by a second face-to-face interviewer visit.

During the first visit, the interviewer collects data on income over the last 12 months, household characteristics (sociodemographic information, characteristics of the dwelling, and qualitative questions about the household's financial situation), and major or recurring expenditures. During this visit, subsample 1 will be provided with the PAPI diary, and subsample 2 will be asked to choose either the PAPI diary or app-based diary, and is given credentials to fill in the web-based questionnaire. All household members over 14 years of age are asked to record their expenditures over seven days (that can be done in a single diary).

During the interviewer's second visit, the paper diaries will be collected and checked for accuracy. If the web questionnaire has been completed, the correspondent modules are skipped in the second interview. The set of questions on the topic of consumption are shortened during the second visit, to accommodate instead a methodological questionnaire. Respondents are asked about their difficulties in filling in their activities in both modes, and about their preferred mode. The interviewer also must fill in a questionnaire on the respondent's quality and engagement.

Aim: Measuring the impact of the digital versus paper diary, and the mode effect on consumption questions, in a realistic context

Research Questions of HBS

UX/UI impact:

- How can one maximize the accuracy of information provided by the app diary? (e.g., through improvements of the app's design, of the product and store lists...)

Acceptability, motivation, quality:

- What mode do respondents prefer, and how does this impact their diary completion?

Appendix D: country documentation for Germany

Survey: HBS, field period: 2 weeks (approximately in September 2024)

Sample: Netto sample of 7.000, two stage random sample (first stage: municipalties, second stage: individuals), target population: German population in the age of 18 to 70.

Materials:

Invitations: (see next page)

UNIVERSITÄT MANNHEIM
Fakultät für Sozialwissenschaften

Prof. Dr. Florian Keusch
Lehrstuhl für Social Data Science & sozialwissenschaftliche Methodenlehre
A5, 6
68131 Mannheim, Germany

Bei Fragen zur Studie können Sie sich an uns wenden:
AusgabenAtlas
ausgabenatlas@uni-mannheim.de
0621 181-2203
https://www.uni-mannheim.de/ausgabenatlas

Universität Mannheim, LS Keusch - A5, 6 · 68131 Mannheim

Michael Mustermann
Musterstrasse 5
345678 Musterstadt

Mannheim, 15. Mai 2024

Wie viel Geld geben die Menschen in Deutschland aus? Helfen Sie uns, es herauszufinden!

Guten Tag Michael Mustermann,

die Lebenshaltungskosten in Deutschland steigen stetig. Doch wie viel Geld geben Bürgerinnen und Bürger wofür aus? Der **AusgabenAtlas** möchte diese Frage wissenschaftlich untersuchen.

Ihr Beitrag zu dieser Studie zählt! Als zufällig ausgewählte Person helfen Sie mit Ihrer Teilnahme ein detailliertes Bild der Ausgaben in Deutschland zu erstellen.

Dafür bitten wir Sie, für einen Zeitraum von zwei Wochen **genaue Angaben zu allen von Ihnen gekauften Produkten und Dienstleistungen in der AusgabenAtlas App zu erfassen.**

Als Dankeschön erhalten Sie einen **10€ Amazon-Gutschein** nach Abschluss der Studie. Ihre Daten behandeln wir vertraulich gemäß den datenschutzrechtlichen Vorgaben.

So einfach machen Sie mit

1. **App "AusgabenAtlas" herunterladen**
   - QR-Code scannen
   - Oder im AppStore nach "AusgabenAtlas" suchen
2. **Mit persönlichen Zugangsdaten anmelden**

Kennung:	823479
Passwort:	Saisons66,Halben

3. **Ausgaben in der App erfassen**

Wir danken für Ihre wertvolle Unterstützung!

Prof. Dr. Florian Keusch

Weitere Informationen auf der Rückseite – bitte wenden.

UNIVERSITÄT MANNHEIM
Fakultät für Sozialwissenschaften

AUSGABENATLAS

Weitere Informationen zur Studie

- Die Universität Mannheim führt die Studie „AusgabenAtlas“ im Rahmen eines Forschungsprojektes durch.
- Ziel ist es, das Konsumverhalten und die monatlichen Ausgaben der Menschen in Deutschland zu analysieren. Solche Daten dienen als wissenschaftliche Grundlage für die Entwicklung einer gerechten Sozial- und Familienpolitik und ermöglichen den Vergleich mit anderen EU-Ländern.

Warum gerade ich?

- Ihre Adresse wurde zufällig von einem Einwohnermeldeamt ausgewählt, um eine repräsentative Stichprobe der Bevölkerung zu gewährleisten.
- Ihre Teilnahme ist entscheidend, um aussagekräftige Ergebnisse über die Lebenssituation in Deutschland zu erhalten, welche die gesamte Bevölkerung widerspiegeln.
- Die Teilnahme ist freiwillig.

Datenschutzgarantie

- Wir garantieren die Einhaltung aller Datenschutzbestimmungen.
- Alle Angaben werden ausschließlich für wissenschaftliche Zwecke erhoben und streng vertraulich behandelt. Diese Studie verfolgt keine gewerblichen Interessen und verpflichtet Sie zu nichts.
- Ausführliche weitere Hinweise zur Studie, sowie Datenschutzhinweise finden Sie unter: https://www.uni-mannheim.de/ausgabenatlas

Kontaktmöglichkeiten

Bei Fragen erreichen Sie uns

- per E-Mail unter ausgabenatlas@uni-mannheim.de
- telefonisch unter 0621 181-2203

Weitere Informationen finden Sie auf unserer Webseite
https://www.uni-mannheim.de/ausgabenatlas

UNIVERSITÄT MANNHEIM
Fakultät für Sozialwissenschaften

Prof. Dr. Florian Keusch
Lehrstuhl für Social Data Science & sozialwissenschaftliche Methodenlehre
A5, 6
68131 Mannheim, Germany

Bei Fragen zur Studie können Sie sich an uns wenden:
AusgabenAtlas
ausgabenatlas@uni-mannheim.de
0621 181-2203
https://www.uni-mannheim.de/ausgabenatlas

Universität Mannheim, LS Keusch - A5, 6 · 68131 Mannheim

Michael Mustermann
Musterstrasse 5
345678 Musterstadt

Mannheim, 15. Mai 2024

Wie viel Geld geben die Menschen in Deutschland aus? Helfen Sie uns, es herauszufinden!

Guten Tag Michael Mustermann,

die Lebenshaltungskosten in Deutschland steigen stetig. Doch wie viel Geld geben Bürgerinnen und Bürger wofür aus? Der **AusgabenAtlas** möchte diese Frage wissenschaftlich untersuchen.

Ihr Beitrag zu dieser Studie zählt! Als zufällig ausgewählte Person helfen Sie mit Ihrer Teilnahme ein detailliertes Bild der Ausgaben in Deutschland zu erstellen.

Dafür bitten wir Sie, für einen Zeitraum von zwei Wochen **genaue Angaben zu allen von Ihnen gekauften Produkten und Dienstleistungen in der AusgabenAtlas App zu erfassen.** Dazu können Sie Ihre **Kassenzettel in der App bequem abfotografieren**.

Als Dankeschön erhalten Sie einen **10€ Amazon-Gutschein** nach Abschluss der Studie. Ihre Daten behandeln wir vertraulich gemäß den datenschutzrechtlichen Vorgaben.

So einfach machen Sie mit

1. **App "AusgabenAtlas" herunterladen**
   - QR-Code scannen
   - Oder im AppStore nach "AusgabenAtlas" suchen
2. **Mit persönlichen Zugangsdaten anmelden**

Kennung:	823479
Passwort:	Saisons66.Halben

3. **Ausgaben in der App erfassen**

Wir danken für Ihre wertvolle Unterstützung!

Prof. Dr. Florian Keusch

Weitere Informationen auf der Rückseite – bitte wenden.

UNIVERSITÄT MANNHEIM
Fakultät für Sozialwissenschaften

AUSGABENATLAS

Weitere Informationen zur Studie

- Die Universität Mannheim führt die Studie „AusgabenAtlas“ im Rahmen eines Forschungsprojektes durch.
- Ziel ist es, das Konsumverhalten und die monatlichen Ausgaben der Menschen in Deutschland zu analysieren. Solche Daten dienen als wissenschaftliche Grundlage für die Entwicklung einer gerechten Sozial- und Familienpolitik und ermöglichen den Vergleich mit anderen EU-Ländern.

Warum gerade ich?

- Ihre Adresse wurde zufällig von einem Einwohnermeldeamt ausgewählt, um eine repräsentative Stichprobe der Bevölkerung zu gewährleisten.
- Ihre Teilnahme ist entscheidend, um aussagekräftige Ergebnisse über die Lebenssituation in Deutschland zu erhalten, welche die gesamte Bevölkerung widerspiegeln.
- Die Teilnahme ist freiwillig.

Datenschutzgarantie

- Wir garantieren die Einhaltung aller Datenschutzbestimmungen.
- Alle Angaben werden ausschließlich für wissenschaftliche Zwecke erhoben und streng vertraulich behandelt. Diese Studie verfolgt keine gewerblichen Interessen und verpflichtet Sie zu nichts.
- Ausführliche weitere Hinweise zur Studie, sowie Datenschutzhinweise finden Sie unter: https://www.uni-mannheim.de/ausgabenatlas

Kontaktmöglichkeiten

Bei Fragen erreichen Sie uns

- per E-Mail unter ausgabenatlas@uni-mannheim.de
- telefonisch unter 0621 181-2203

Weitere Informationen finden Sie auf unserer Webseite
https://www.uni-mannheim.de/ausgabenatlas

UNIVERSITÄT MANNHEIM
Fakultät für Sozialwissenschaften

Prof. Dr. Florian Keusch
Lehrstuhl für Social Data Science & sozialwissenschaftliche Methodenlehre
A5, 6
68131 Mannheim, Germany

Bei Fragen zur Studie können Sie sich an uns wenden:
AusgabenAtlas
ausgabenatlas@uni-mannheim.de
0621 181-2203
https://www.uni-mannheim.de/ausgabenatlas

Universität Mannheim, LS Keusch - A5, 6 · 68131 Mannheim

Michael Mustermann
Musterstrasse 5
345678 Musterstadt

Mannheim, 15. Mai 2024

Wie viel Geld geben die Menschen in Deutschland aus? Helfen Sie uns, es herauszufinden!

Guten Tag Michael Mustermann,

die Lebenshaltungskosten in Deutschland steigen stetig. Doch wie viel Geld geben Bürgerinnen und Bürger wofür aus? Der **AusgabenAtlas** möchte diese Frage wissenschaftlich untersuchen.

Ihr Beitrag zu dieser Studie zählt! Als zufällig ausgewählte Person helfen Sie mit Ihrer Teilnahme ein detailliertes Bild der Ausgaben in Deutschland zu erstellen.

Dafür bitten wir Sie, für einen Zeitraum von zwei Wochen **genaue Angaben zu allen von Ihnen gekauften Produkten und Dienstleistungen in der AusgabenAtlas App zu erfassen.** Dies beinhaltet für jedes gekaufte Produkt den Preis, die Menge und die Produktkategorie. Dazu können Sie Ihre **Kassenzettel in der App bequem abfotografieren**.

Als Dankeschön erhalten Sie einen **10€ Amazon-Gutschein** nach Abschluss der Studie. Ihre Daten behandeln wir vertraulich gemäß den datenschutzrechtlichen Vorgaben.

So einfach machen Sie mit

1. **App "AusgabenAtlas" herunterladen**
   - QR-Code scannen
   - Oder im AppStore nach "AusgabenAtlas" suchen
2. **Mit persönlichen Zugangsdaten anmelden**

Kennung:	823479
Passwort:	Saisons66.Halben

3. **Ausgaben in der App erfassen**

Wir danken für Ihre wertvolle Unterstützung!

Prof. Dr. Florian Keusch

Weitere Informationen auf der Rückseite – bitte wenden.

UNIVERSITÄT MANNHEIM
Fakultät für Sozialwissenschaften

AUSGABENATLAS

Weitere Informationen zur Studie

- Die Universität Mannheim führt die Studie „AusgabenAtlas“ im Rahmen eines Forschungsprojektes durch.
- Ziel ist es, das Konsumverhalten und die monatlichen Ausgaben der Menschen in Deutschland zu analysieren. Solche Daten dienen als wissenschaftliche Grundlage für die Entwicklung einer gerechten Sozial- und Familienpolitik und ermöglichen den Vergleich mit anderen EU-Ländern.

Warum gerade ich?

- Ihre Adresse wurde zufällig von einem Einwohnermeldeamt ausgewählt, um eine repräsentative Stichprobe der Bevölkerung zu gewährleisten.
- Ihre Teilnahme ist entscheidend, um aussagekräftige Ergebnisse über die Lebenssituation in Deutschland zu erhalten, welche die gesamte Bevölkerung widerspiegeln.
- Die Teilnahme ist freiwillig.

Datenschutzgarantie

- Wir garantieren die Einhaltung aller Datenschutzbestimmungen.
- Alle Angaben werden ausschließlich für wissenschaftliche Zwecke erhoben und streng vertraulich behandelt. Diese Studie verfolgt keine gewerblichen Interessen und verpflichtet Sie zu nichts.
- Ausführliche weitere Hinweise zur Studie, sowie Datenschutzhinweise finden Sie unter: https://www.uni-mannheim.de/ausgabenatlas

Kontaktmöglichkeiten

Bei Fragen erreichen Sie uns

- per E-Mail unter ausgabenatlas@uni-mannheim.de
- telefonisch unter 0621 181-2203

Weitere Informationen finden Sie auf unserer Webseite
https://www.uni-mannheim.de/ausgabenatlas

2

Appendix E: country documentation for Italy

Methodological Test of TUS by ISTAT

Date: October/November 2024 *(more details to follow as materials are prepared in next version of deliverable)*

Net sample: 5,000 households

Modes: app-based TUS

Incentive structure:

Protocol

Italy will deploy interviewers to each household for recruitment and assistance purposes. Each selected household will receive a pre-notification letter via postal mail. This letter will inform them of their selection for the study and provide details about the upcoming interviewer visit. The sampled households will be randomly assigned to one of two experimental conditions: high-smart and low-smart, which will alter the function of the app accordingly.

Detailed protocol information:

Aim: Testing the added benefit of a high-smart feature and assessing potential consequences for differences in measurement.

Experimental conditions

The high-smart condition will sample users' geolocations within the MOTUS app. Microservices on the server side will segment the timestamped location data into trips and stops, deriving time use entries related to user travel behavior. The low-smart condition will not record location behavior or prepopulate certain fields.

MOTUS

Italy will use the MOTUS app for time use registration. Interviewers will help respondents install the app at the first visit.

Additional variables

Research questions

- Does prepopulating the diary using geolocations impact the content of the TUS diary?
- Does prepopulating the diary using geolocations impact measures of quality of the TUS diary?

Appendix F: country documentation for the Netherlands

Methodological test of HBS by CBS

Date: September/October 2024 *(more details to follow in new version of deliverable when materials are completed)*

Net sample: 1,600 households residing in the Netherlands, selected from the population register, subdivided across three different subsamples on the basis of predicted reachability, and allocated to one of three levels of interviewer involvement.

Modes: App-based diary

Incentive structure: Unconditional incentive

Protocol

1,600 households are contacted by postal mail at their registered address with a letter, prenotification and letter, or prenotification and interviewer, depending on experimental condition. Respondents complete an initial survey on household characteristics. Following this, respondents are requested to record all purchase details for a two-week specified reporting period.

Detailed protocol information

Aim: To investigate the effectiveness of using interviewers and prenotification letters to boost participation, especially among hard-to-reach groups.

Experimental conditions

Experiment one (*Reachability*) involves the selection of three distinct subsamples: 1) a subsample that mirrors the target population; 2) a subsample consisting of hard-to-reach respondents (selected on the basis of age and origin); and 3) a subsample composed of individuals with high participation likelihood (selected on the basis of age and origin).

Experiment two (*Interviewer involvement*) varies the contact method. Sampled individuals are randomly assigned to one of three conditions: 1) receiving an invitation letter only, 2) receiving a prenotification letter before receiving the invitation letter, and 3) receiving the invitation letter, followed by an interviewer visit for recruitment and assistance with the app.

Research Questions

- Does the usage of interviewers during recruitment positively affect response rate?
- Does the usage of announcement letters during recruitment positively affect response rates?
- Does the usage of interviewers and announcement letters help to reach respondents with low participation probabilities?

Appendix G: country documentation for Norway

SSI/WP2.1: Larger Recruitment Field Test - Statistics Norway (SSB)

This appendix describes the test design at Statistics Norway (SSB)

Test design

Survey instrument

Statistics Norway (SSB) have developed a survey design to assess the impact of employing CATI interviewers or not, and two different methods of login to survey. We will use the setup and data collection instrument for the Household Budget Survey (HBS) 2022 in Norway. The survey instrument is the Progressive Web App (PWA) or web app we used in 2022, which means it can be opened on any kind of device with an internet connection and used both online and offline. The diary will be identical with that of HBS 2022, offering options between OCR and manual registration of purchases and products items bought. The questionnaire section will be somewhat reduced (also included in the PWA). Household composition data will be uploaded from the national register and not confirmed in CATI interviews like for HBS 2022.

Sample

A random gross sample of 2,000 households will be drawn from the national register. This is the official government register that is also used by the tax office. This register is updated continuously and reach almost all citizens. Messages sent to e-mail addresses and mobile numbers for this register are read shortly after delivery by a large majority of respondents.

Experiment

The sample will be split in half: 1) Group one will receive CATI recruitment and follow-up, while 2) Group 2 will have not interviewer recruitment or follow up. Each sub-sample will further be split in half: A) One subgroup/low trust receiving a SMS with web link to the app and B) One subgroup/high trust receiving a SMS that ask respondents to go to Altinn's homepage. Altinn is a government service or platform for communication with citizens. It is used for the tax form and other official forms. Group 1A with CATI and sms link directly to app is comparable to actual set up for HBS 2022.

Altinn has a two-step-authentication login solution with ID-porten/Bank-ID which is used for all government, bank, health services etc. in Norway. Upon first login, respondents have to login with ID-porten/Bank_ID to Altinn and then again to get access to the webapp. ID-porten/Bank-ID will not be asked again for later logins.

Table G1: Test design, groups, and sample size:

Recruitment experiment: N=2 000	Trust	
	A) **LOW** trust sms with link to app & ID-porten	B) HIGH trust sms: "Go to homepage for Altinn"
1) WITH CATI	500	500
2) WITHOUT CATI	500	500

Dispatch plan

- Sample size: 2 000 units gross sample (household sample, contact person selected)
- Field period: Week 14-20, 2024
- Week 14: We start recruitment of the first batch with a sample of 1 000
    - Half is recruited by CATI interviewers (start Monday) and half without (dispatch Friday)
    - Each half are divided in login with link and login with homepage
- Week 15: We send out login info to batch one
    - This is the first reference week of batch one
- Week 16: We start recruitment of the second batch with an additional sample of 1 000
    - Same set up as for batch one
- Week 17: We send out login info to batch two
    - This is the first reference week for batch two

- Change of reference week is possible for all within a 2-week period.
    - Contact or reference person is set in agreement with respondents during recruitment interview for group with CATI. For group without CATI, SSB preselects.

Table G2: Dispatch plan

Week	Sample	Sub sample	CATI recruitment	Login with	n=
Batch one					
W14	1000	500	Yes	sms	250
				Altinn	250
		500	No	link	250
				Altinn	250
W15	Registration, follow-ups, and some refusal-follow-ups (recruitment)				
Batch two					
W16	1000	500	Yes	sms	250
				Altinn	250
		500	No	sms	250
				Altinn	250
W17	Registration, follow-ups, and some refusal-follow-ups (recruitment)				
W18	Registration, follow-ups, and some refusal-follow-ups				
W19	Registration, follow-ups, and some refusal-follow-ups				
W20	Registration, follow-ups, and some refusal-follow-ups				
SUM:					2000

Contact plan and survey communication

Selected sample for all groups will all receive an e-mail with information that they are selected to participate in the survey. Participation is voluntary.

After this the groups with and without CATI interviewers receive slightly different contact plans. This is illustrated in separate figures below:

Table G3: Survey communication – Without CATI interviewers

Week	Weekday	Contact type	Category	Interview status	Age
W-1	Thursday	Altinn	Information letter about survey and selected reference week	Selected sample (Reference person and week selected by SSB*)	All
W0	Monday	Altinn	Letter in Altinn with login info	Not Started	All
		SMS	SMS with login information with to homepage	Not Started	All
	Wednesday	SMS	Tips for started respondents	Started	All
		SMS	SMS those who have not started	Not started	All
	Thursday	SMS	Text those who have few registrations	Started	All
	Friday	SMS	Reminder regarding weekend expenses	Started	All
		SMS	SMS those who have not started with information about pushing the registration week.	Not started	All
W+1	Monday	Email	Thank you email	Finished	All

*The reference week will be possible to change in app for this group.

Note that the groups without CATI interviewers will not receive any interviewer contact, neither CATI recruitment, nor CATI follow-up.

Survey communication – With CATI interviewers

The contact plan for groups with CATI interviewers is similar but text/content of communication is slightly different.

Table G4: Survey communication with interviewers

Week	Weekday	Contact type	Category	Recruitment status	Age
W-2	Thursday	Altinn	Information letter about survey, and CATI contact from SSB	Selected sample (Reference person and week in agreement w/HH*)	All
W-1	Monday	SMS	SMS about CATI recruitment	Selected sample	All
	Wednesday	SMS	Reminder	Not recruited	All
	Thursday	SMS	Reminder	Not recruited	All
W0	Monday	Altinn	Email with login information	Recruited	65-79 years
		SMS	SMS with login information	Recruited	All
	Thursday	CATI	CATI to those who have not started	Not started	All
	Wednesday	SMS	Tips for started respondents	Started	All
	Thursday	CATI	Call those who have not started	Not started	All
		CATI	CATI those who have few registrations	Started	All
	Friday	SMS	Reminder regarding weekend expenses	Started	All
W+1	Monday	Email	Thank you email	Finished	All

*The reference week will be possible to change in CATI recruitment interview within a 2-week period.

Incentives

Conditional incentives upon completion to all participants. They will receive a gift certificate to the value of NOK 500.

Timing

- Project planning – biweekly meetings from sept 2023
- Set up – December -23 – February -24
    - Revised questionnaire ready for programming – Friday 9.2.24
    - Programmed questionnaire/app finished – Tuesday 20.2.24
    - Micro service for log in solution to be added to PWA – Tuesday 20.2.24
- Field – April and May week 14-20, 2024
    - Start recruitment – Tuesday 2. 4
    - First reference week 15 – Monday 8. 4
    - End of field – Sunday 19. 5
- Data – End of June 2024 – Friday 28.6
- Document report – September 2024

Data delivery

All data to be sent to researchers abroad has to be anonymous. Data will mainly be process data, but also some survey data. Data delivery is described in a separate Excel file. There is one sheet for the net sample and one for gross sample.

Data agreement

SSB and the consortium has an overall agreement about the project and SSB's contribution. SSB does not require a separate data handler agreement in edition for the data delivery, as long data is anonymous.

Note that SSB expects data to be used only for the ESS-SSI 2023-25 project only; and that the data will be analysed and handled by the University of Mannheim. When statistics is published SSB expects to be credited as data supplier.