|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | Logo STATEC - Institut National de la Statistique et des Etudes Economiques | Logo von Statistik Austria |
|  |  |  |  |  |  |



****

**ESSnet Smart Surveys**

**Grant Agreement Number: 899365 - 2019-DE-SmartStat**

[Link to our CROS website](https://ec.europa.eu/eurostat/cros/content/essnet-smart-surveys_en)

**Workpackage 2**

**Smart Survey Pilots**

**Deliverable 2.4: The role of interviewers**

**Version 1.2, 31-05-2022**

**Prepared by:**

Jannick Akkermans, Evelien Rodenburg, Jelmer de Groot, Barry Schouten (CBS, Netherlands)

Encarnacion Martin Bernia, Victor Balsa Criado, Tamara Horcajo (INE, Spain)

Catherine Gauche, Guillaume Osier (STATEC, Luxemburg)

ESSnet co-ordinator:

Workpackage Leader:

Barry Schouten (CBS, Netherlands)

[Jg.schouten@cbs.nl](mailto:Jg.schouten@cbs.nl)

telephone: +31 70 3374905

SUMMARY: WP2 performs four diverse pilots to inform WP3 on the specifications of a smart survey platform in the European Statistical System (ESS). An influential decision is the employment of interviewers in recruitment and motivation of respondents. In phase 3 of the Household Budget Survey pilot interviewer-assistance was randomized in the samples of the three participating countries (ES, LU and NL). This deliverable reports the impact of interviewer-assistance on recruitment, in-app activity and HBS data quality.

1. INTRODUCTION

This deliverable is one of three reports in ESSnet Smart Surveys evaluating the application of the Household Budget Survey app in a large-scale field test.

The Household Budget Survey app has been developed within Eurostat-funded projects @HBS (Jan 2019 to February 2020) and @HBS2 (March 2021 to April 2022). In ESSnet Smart Surveys the app has been modified in two stages to six new countries: BE, DE, ES, LU, NO, and PL. The app already existed in country versions for FI, NL, SI and UK prior to the ESSnet, but several revisions have been made and extensions added during the ESSnet. The six new countries performed functional tests within the ESSnet during May 2020 to December 2020. Revisions of the app, based on these phase 2 tests, are reported in deliverable 2.9. on WP2 specifications for WP3. Country differences and commonalities are reported in deliverable 2.5 on shareability. Three countries, ES, LU and NL, participated in the phase 3 field test and for these countries the app went through another modification stage between Jan 2021 and September 2021.

Underlying to the phase 3 HBS field test are the following questions as listed in the ESSnet project proposal:

* How does an app-assisted HBS affect recruitment rates?
* How does an app-assisted HBS affect drop-out/completion rates?
* How does an app-assisted HBS affect data quality?
* How does an app-assisted HBS affect substantive plausibility?

These four questions are discussed in general in deliverable 2.1, elaborated for respondent statistics feedback and interaction in deliverable 2.3 and for the role of interviewers in this deliverable 2.4. In deliverable 2.1a, the HBS statistics based on the field test data are evaluated in terms of plausibility and compared to the regular, ‘non-smart’ HBS. In this deliverable, only relative comparisons are made between the interviewer and non-interviewer samples.

So what motivates the interest in interviewers in an app-assisted HBS? To discuss this role, a step back is needed. Smart surveys aim to reduce respondent burden, automate certain respondent tasks and improve survey/user experience. They can do so through a range of smart features: 1) in-device storage and processing, 2) application of internal mobile device sensors, 3) linkage to external sensor systems, 4) linkage to public online data, 5) data donation through the respondent and 6) data donation through the statistical institute. However, sensor data, donated data and other forms of linked data are not free from error. Respondents can be ask to assist in correcting errors or providing context. Such respondent-interaction will introduce a new task and may thus partly undo the perceived reduction of burden through smart features. Such tasks may also be complex to some as they involve correctly navigating user interfaces and handling devices adequately. The interviewer may play a prominent role in this respondent active-passive involvement trade-off in the following roles:

1. Explain the improved survey experience;
2. Explain the respondent task and remove any hesitation;
3. Assist in starting/installing the app;
4. Become an intermediary between the survey institute and the respondent during the data collection period;
5. Motivate respondents more generally of the purpose and utility of the survey;

The importance of these roles differs between smart surveys. The last role is the most general; interviewers always increase recruitment rates and completion rates. They personalize communication and show the value that the survey institute attaches to participation. There is a vast literature on the role of interviewers that is not revisited here. The other four roles are much more specific to smart surveys. Respondents may identify themselves with interviewers and interviewer perceptions on the user experience and complexity of the respondent task can be convincing. Having had training in the smart survey tools, interviewers may also help out and even become a first starting point for help during data collection. Obviously, interviewers are no mobile app developers, UI/UX designers, or data scientists, so that their assistance must be pragmatic and other channels are needed for detailed assistance.

Getting back to the HBS, the role of interviewers may be in all five roles described above. They can explain that the app provides personalized feedback, they can give information on the time and receipt scanning option, they can show the app and can give some guidance on how to install and use it.

An important follow-up question to the utility of interviewer-assistance is how the impact of interviewers can be evaluated? In the field test, interviewer-assistance was randomized; one sample had assistance and the other not. The non-interviewer sample thus is a control group. To quantify differences, three questions are asked:

1. what is the impact of interviewer-assistance on registration and completion rates?
2. what is the impact of interviewer-assistance on data quality?
3. what is the recommended role of interviewers?

The impact on data quality is the most complex as true values are unknown. Data quality will be substantiated in this deliverable.

The deliverable has the following outline. Section 2 elaborates the metrics used to evaluate interviewer impact, in particular data quality. Section 3 gives details on the experimental design and interviewer training/instruction. Then in section 4, overall results are presented. Sections 5 and 6 consider, respectively, recruitment, activity and completion (RQ1) and data quality (RQ2). The deliverable is ended with a discussion of the recommended role of interviewers (RQ3) and next steps in section 7.

1. METHODOLOGY

In quantifying evaluations, the two main dimensions recruitment (representation) and data quality (measurement) are considered. For both dimensions, different data can be employed to evaluate.

* 1. RECRUITMENT AND COMPLETION

* + 1. RATES

If an interviewer is preferred as mode of approach, it is important to look at the recruitment rates and completion rates. It is not just in the registration rates where the added value of an interviewer can be seen, but also if an interviewer can recruit within different age groups or household sizes that are less responsive to a letter-only approach. It could be that some persons are more likely to participate in the survey when they are being invited by an interviewer at the door versus when they are being approached by only a letter with log-in codes.

The different indicators that are measured here to investigate this, are:

* Overall response rates & comparison between different age groups
* Overall completion rates & comparison between different age groups

2.1.2 RECRUITMENT AND COMPLETION: INTERVIEWER PERCEPTIONS

A lot of the behaviour in the app can be measured through the data that respondents reported in the app. However, these data do not show how the interviewers’ visits at the door went, what interviewers encountered, etc. The interviewers are our only window to the respondents’ minds. The interviewers talked to the respodents at the door and they also did a telephone motivational call. These two contact moments can tell much about the role of the interviewer.

What do interviewers themselves say about the visits they did at the door? Were there any additional things they had to say to convince people to take part? What did people tell to the interviewers when they had a motivational call during the fieldwork period?

What interviewers have experienced at the door can be of great use for further exploration and implementation of a data collection strategy.

* 1. DATA QUALITY
     1. RICHNESS OF REPORTED HBS DATA

Having someone at the door might have a positive effect of the motivation and involvement of the respondent. Because of the personal ‘touch’ of the interviewer people might get higher motivated to report their purchases. This can be seen in de amount of data households are reporting. In the data that was analysed it could also be made visible if different indicators for data quality declined over time, for example when people are losing motivation over time. This loss of motivation can be compared between the interviewer versus non-interviewer group.

It is not straightforward how data quality in the HBS survey can be measured. Intuitively it makes sense that ‘good’ quality expenditure data should be more diverse, i.e., both small and large purchases are reported, and purchases are reported in multiple store types. Also, the more purchases people report in the app, the higher the quality of the data could be. Thus, the following seven indicators are used to get a sense of the quality of the expenditure data:

* the number of purchases per household (entries);
* the average amount of the entries
* the difference between the maximum and minimum amount spent per entry per household (amount variation);
* the standard deviation of the amount of money spent per entry within a household (SD amount);
* the difference between the maximum and minimum number of products bought per entry per household (products variation);
* the standard deviation of the number of products bought per entry within a household (SD products);
* the expected number of different store types in which a household entered a purchase (store types).

In order to know if the data is of enough quality, the above data quality indicators were compared with the AU week (all purchases) week of the regular HBS survey.

The quality indicators will be calculated separately for week 1 (days 1-7), week 2 (days 8-14), and for the whole study period (overall). Because households were still able to use the app after the study period ended, the overall indicators can also contain information on expenditures entered after the second week. The indicators will also be calculated for the AU (all expenses) week of the 2020 HBS. To standardize the amounts and numbers of products and eliminate country effects, the amount per household is divided by the average amount per country, and the number of products per household is divided by the average number of products per country. This is done separately for week 1, week 2, overall, and for the 2020 data (using the inclusion weights to obtain weighted averages). The standardized amounts and numbers of products will be used to calculate the amount variation, SD amount, product variation, and SD product.

Because the intent is to measure the variation in the data, the amount variation, SD amount, product variation, and SD product are set to missing for households with only one entry. This is again done separately for week 1, week 2, overall, and for the 2020 data.

Finally, longer study durations give households more opportunities to report purchases in more store types. To correct for the difference in study durations, an expected number of store types is calculated using the following formula:

),

where represents the store type and represents a household. The probability of a purchase entered by a household being a certain store type is represented by , and represents the number of purchases entered by a household. To standardize the number of store types, the can be set to a fixed value, where the are estimated based on reported data.

The data that was included within these quality indicators are the 2020 sample (AU week) and for the HBS 2021 sample that filled out week 1. This is done because not all countries have sufficient data for the second week. For 2021, the above seven indicators are calculated for all households that provided at least one purchase, and for the regular HBS 2020 the seven indicators are calculated for all households that completed the survey. The inclusion weights will be used in the calculations of the quality indicators for the 2020 HBS data to correct for the differences in the sampling designs between 2020 and 2021. Comparisons between the data quality in 2021 and 2020 will be done by comparing means (including confidence intervals), quartiles, and standard deviations.

2.2.2 IN-APP ANSWERING BEHAVIOUR

Next to the analyses that are done for the data quality, another in-depth analysis is made for the app behavior that respondents showed. There were different conditions used (interviewer yes/no) to see if this does something with the motivation and involvement of the respondents. It can be argued that more motivated respondents report more purchases per day and that these entries have higher amounts. The more people are involved and motivated, the more entries they put into the app and the more precise this is done. So it could be argued that this can be seen in the number of entries as well as the amounts of purchase entered. To be able to say something about the involvement and motivation of respondents, different types of data could be analyzed.

This was done for the following indicators:

* The number of entries per day, and split into manual/scanned entries
* The amount of the these entries per household, also split into manual/ scanned entries

2.2.3 Paradata analyses

Besides the already mentioned data quality indicators, paradata also led to extra richness in data analysis. It sounds logic to think that people whore are more motivated and involved in using the app, will report more purchases and provide data that is of higher quality. This paradata provides meaningful information about the respondents’ app behavior. Motivated respondents might use more screens and spend more screen time than less motivated respondents. It might be the case that respondents In order to analyze this, the following indicators were analyzed:

* The time spent in the app per day
* Number of app pages used (average & total) per day

To investigate the best approach strategy for app surveys, these quality indicators were compared between the interviewer versus non-interviewer group to see what the benefits and trade-offs can be of an interviewer assisted approach. Also, the above indicators for both involvement in the app on the basis of purchases and on the analysis of paradata can be linked for even more richer analyses.

1. DESIGN OF THE HBS FIELD TEST

3.1 THE HBS APP

The phase 3 field test employed version 2.1.15 of the Household Budget Survey app available in Apple and Google app stores. The app could be used on smart phones and on tablets in portrait mode, but not on desktops and laptops. Technical performance was guaranteed for operating systems from 2016 on. The app version 2.1.15 included a number of user interface revisions suggested by functional/usability phase 2 tests within ESSnet Smart Surveys as well as a number of extensions prepared within Eurostat-funded project @HBS2. The functional/usability tests were accompanied by a short questionnaire to all WP2.1 countries (BE, DE, ES, LU, NL, NO, PL) on the desired frontend and backend specifications.

The revised app features are:

* In-app tutorial movie to instruct respondents on how to scan receipts
* Adjustment of parameters in the product search algorithm (using a Jaro-Winkler distance function)
* Revision of discount and rebate data entry options
* Option to turn on daily reminders
* Textual changes in some of the UI screens

The new app features are:

* Automatic receipt scan detection
* The option to crop receipt scan images
* In-app text recognition including editing options by respondents
* Inclusion of paradata on in-app navigation behaviour
* A configurable in-app intro questionnaire

During the preparations for the field test, it was decided not to display in-house text recognition and COICOP classification results in-app to respondents. The main reasons were the complex backend database structure that would be needed to keep track of unchecked, checked and revised receipts and the burdensome and complicated additional respondent classification task.

3.2 SAMPLE AND EXPERIMENTAL DESIGN

For various reasons, BE, DE, NO and PL decided not to participate in the phase 3 field test. The anticipated sample size of 4000 households was, therefore, allocated to the three remaining countries, ES, LU and NL. The original anticipated sample size per country was 800, assuming five participating countries. LU and NL agreed to double their sample sizes in order to meet statistical power requirements. The resulting sample sizes were 1600 for LU and NL and 800 for ES. Ultimately the sample size for LU was slightly larger, because the amount of administrative nonresponse was smaller than anticipated. The sample size for NL was slightly larger as the workload for face-to-face interviewers had to be slightly reduced because of COVID-19 procedures.

Table 1 presents the sample sizes per country and condition.

*Table 1: Sample sizes per country and experimental condition.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | ES | LU | NL |
| Feedback instant | Planned: 400  Realized: 433 | Planned: 800  Realized: 882 | Planned: 800  Realized: 748 |
| Feedback delayed | Planned: 400  Realized: 433 | Planned: 800  Realized: 884 | Planned: 800  Realized: 737 |
| Interviewer | Planned: 400  Realized: 433 | Planned: 800  Realized: 881 | Planned: 800  Realized: 685 |
| No interviewer | Planned: 400  Realized: 433 | Planned: 800  Realized: 884 | Planned: 800  Realized: 800 |
| In-app editing | Planned: 800  Realized: 866 | - | Planned: 1200  Realized: 1085 |
| No in-app editing | - | Planned: 1600  Realized: 1766 | Planned: 400  Realized: 400 |

The samples were randomly split to the feedback condition. One half got instant feedback. Under this condition, feedback was explicitly mentioned in invitation letters and interviewers could use it as an argument at first contact. The other half got feedback on the last reporting day and no mention was made in invitation letters or by interviewers. For these households the app Insights screen displayed a so-called empty state layout indicating that feedback would only be delivered later during the reporting period.

The samples were also randomly split to the interviewer condition. However, in LU the interviewer condition implied that interviewers were mentioned as contact persons in invitations. Interviewers did not actually visit or call households. The non-interviewer condition in LU meant that no mention was made of interviewers. In ES and NL, one half of the sample received only a written invitation including a brochure and instruction material. The other half was visited at the door by an interviewer and a motivational call was made a few days later. Annex B presents recruitment material for ES, LU and NL.

The in-app condition was only in part randomized, since it was not planned in the original design and was not enforced.. In ES, in-app editing was always enabled, while in LU it was always disabled. In NL, in-app editing was enabled for the interviewer sample and randomized for the non-interviewer sample. The choice not to randomize within the interviewer sample was made in order to not confuse interviewers when interacting with households. While not randomized within countries, the total samples with and without in-app editing were equal size.

The sample designs varied per country. In ES and NL simple random samples were drawn within selected regions of the country. Since around 20 households per interviewer was deemed optimal in terms of training and experience, samples were drawn in specific interviewer regions. These regions were spread across the countries from rural to urban areas. In LU samples were randomly drawn from the SILC panel, however. Since interviewers were not directly involved in recruiting respondents, no clustering in geographical regions was needed. Samples were not stratified based on household characteristics such as age, income or household composition, as is done in some countries such as NL, in order to optimize statistical power in evaluations.

It is important to remark that the data reporting period varies between LU-NL and ES. LU and NL both invited households to report all expenditures for two weeks, while ES asked full reporting only for one week. This choice was made in order to conform to the regular HBS design in ES and LU. In NL, the regular full reporting period is one week, but a two week period is desired; the HBS phase 3 field test thus also implicitly allows for an evaluation of length of the reporting period. The one week/two weeks reporting period started on the day the household registered the app, i.e. logged in for the first time. In the app calendar screen the reporting days were highlighted and a completed day changed colour from blue to green.

Three experimental conditions were included in order to evaluate push-to-app strategies and active-passive data collection trade-offs:

* Yes/no instant feedback of individual household expenditure statistics
* Yes/no interviewer-assistance in recruiting and motivating respondents
* Yes/no in-app respondent editing of text recognition; this feature is discussed in detail in deliverable 2.3.

The three conditions were randomized independently, i.e. the potential interaction between the conditions was not deemed influential enough to evaluate and is ignored. It would also imply a larger sample size to reach acceptable statistical power The in-app edit option implied that only the edited data was submitted to the backend. However, through paradata logging it can be deduced afterwards whether respondents actually performed editing. In all cases the scanned receipts were submitted to the backend, i.e. regardless of in-app editing. To safeguard privacy, respondents were instructed to crop images to the product-price section on receipts.

3.3 INTERVIEWER PILOT AND INSTRUCTIONS

The sample was divided into two randomly assigned conditions. One condition were persons that were visited by an interviewer and one condition where only an invitation letter was sent. All interviewers where given the opportunity to use the app for a week before they participated in the instruction session, so they were able to ask the questions about the app they have for themselves. Problems that they encountered with the app could be talked about in one of the instruction sessions. In all countries, instructions sessions were held via video meetings by giving PowerPoint presentations by the instructors. These instruction sessions were always followed by an option for interviewers to raise questions they had. Interviewers already were trained in how to react to participants when they open the door. The instructions that were given additionally were HBS-specific. So after the week of use for the interviewers, features of the app could be explained in more detail and the purpose of the research was explained. A paper manual was sent to the interviewers with background information about the HBS survey, installation information for the app and also a FAQ section. Also, interviewers received the leaflet that was used to approach the addresses.

1. FIELDWORK RESULTS IN GENERAL

In general, the context for the HBS phase 3 field test was unfavourable. In all countries, COVID-19 measures were in place and it is was uncertain at the start whether the face-to-face part would be affected by restrictions at the door. The fieldwork was planned from September 1st to November 30th in NL within two batches of households with a month time lag. In ES and LU, fieldwork was planned from November 1st to December 31st. LU released the sample in a single step, but applied several reminders. ES fielded the sample in four batches with one week time lags. In NL, the face-to-face sample faced restrictions from the second half of October onwards to the end of data collection. For this reason, 125 of the face-to-face households were never fielded and are left out of the analysis. In ES, face-to-face fieldwork had to be stopped towards early December in most of the regions. A total of 143 of the households that were in the field were moved to telephone recruitment. In LU, there were no restrictions as interviewers were only employed as contact persons.

The incentive strategy varied per country. In ES, no incentive was given to respondents. In NL, an unconditional incentive of 5 Euro was sent/given at the invitation and a conditional incentive of 20 Euro was given upon completion. In LU, the incentive was dependent on the size of the household and ranged from 40 Euro for a single person household to 60 Euro for a household with children.

Table 2 provides the recruitment, activity and completion rates in the three countries. Activity is defined as submitting at least one purchase. Completion is defined as showing in-app activity from on the last day of data collection. Strictly speaking, households can thus complete the HBS diary without being active. However, this combination occurred only a few times in each country. For ES, rates are given with and without the households that were moved to telephone recruitment.

*Table 2: Registration, activity and completion rates per country per mode condition including standard errors by normal approximation.*

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Registration rate (in %) | Activity rate (in %) | Completion rate (in%) |
| ES | F2F + tel: 25.4 (2.1)  F2F only: 29.4 (2.8)  No-F2F: 11.3 (1.5) | F2F + tel: 21.5 (2.0)  F2F only: 26.5 (2.7)  No-F2F: 10.9 (1.5) | F2F + tel: 18.8 (1.9)  F2F only: 22.4 (2.5)  No-F2F: 8.3 (1.3) |
| LU | F2F contact: 30.1 (1.6)  No F2F contact: 28.0 (1.6) | F2F contact: 22.9 (1.5)  No F2F contact: 21.3 (1.4) | F2F contact: 18.3 (1.4)  No F2F contact: 17.6 (1.3) |
| NL | F2F: 25.6 (1.7)  No-F2F: 15.9 (1.3) | F2F: 23.8 (1.7)  No-F2F: 11.6 (1.1) | F2F: 20.3 (1.0)  No-F2F: 9.6 (1.0) |

Table 2 shows that the three countries have relatively similar registration rates when interviewers are employed and when not. For ES and NL also activity rates and completion rates are quite similar. However, for LU activity rates and completion rates show a stronger decline with respect to registration.

1. RECRUITMENT AND COMPLETION
   1. RECRUITMENT, ACTIVITY AND COMPLETION RATES

From the previous paragraph, it can be conjectured that the contact mode has an effect on the response and activity rates. Table 3 shows the differences between the contact mode and the registration and activity rates.

Table 3 shows that the contact mode causes a significant difference for both registration and activity rates. So It can be concluded that interviewers have a positive effect on response rates. However, it is much more interesting to see the impact on representation of the interviewer based approach. In tables 4 and 5, household size distribution and age distribution are addressed, respectively, to see if there are differences in heterogeneity for both groups.

*Table 3: Difference in registration and activity between the contact mode conditions*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Registered |  |  | Active |  |  |
|  | % |  | *P* | % |  | *P* |
|  |  | 45.35 | <.001 |  | 52.82 | <.001 |
| Interviewer | 25.6 [23.0, 28.3] |  |  | 22.8 [20.3, 25.3] |  |  |
| Mail | 14.3 [12.4, 16.3] |  |  | 11.4 [9.6, 13.2] |  |  |

*Note.* Bootstrap confidence intervals are given in square brackets.

*Table 4: age distribution for both the letter group and the interviewer group*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Sample | Response | % of sample | Letter response | Interviewer response |
| 18-24 | 177 | 29 | 11,3 | 9 (2,2%) | 20 (3,9%) |
| 25-34 | 607 | 159 | 15,0 | 68 (16,7%) | 91 (17,7%) |
| 35-44 | 953 | 263 | 14,8 | 122 (29,9%) | 141 (27,4%) |
| 45-54 | 870 | 189 | 12,0 | 85 (20,8%) | 104 (20,2%) |
| 55-64 | 808 | 139 | 9,8 | 60 (1,47%) | 79 (15,4%) |
| 65-74 | 488 | 81 | 9,0 | 37 (9,1%) | 44 (8,6%) |
| >= 75 | 318 | 62 | 11,0 | 27 (6,6%) | 35 (6,8%) |
|  | 4221 | 922 | 82,9 | 408 | 514 |

*Table 5: household size distribution for both groups.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Class | Sample | Response | Interviewer sample | Interviewer response | Letter sample | Letter response |
| 1 person | 1072 | 195 (18,2%) | 512 | 110 (21,4%) | 560 | 85 (20,8%) |
| 2 persons | 1305 | 293 (22,3%) | 676 | 166 (32,3%) | 629 | 127 (31,1%) |
| 3 persons | 797 | 202 (25,3%) | 407 | 116 (22,6%) | 390 | 86 (21,1%) |
| 4 or more persons | 1047 | 232 (22,2%) | 514 | 122 (23,7%) | 533 | 110 (27%) |

Tables 4 and 5 show the distribution of the age and household size over the pooled respondents for ES, LU and NL. Between the groups with and without interviewers there are no strong differences in age and household size. Both groups seem to be composed of somewhat the same aged people as well as the same types of households. So there is no strong evidence that interviewers recruit specific groups better than when only a letter is sent to these groups. Obviously, age and household size are not the only relevant background characteristics. For NL, a much deeper look into representation would be possible given the rich set of auxiliary variables from administrative data. However, the sample of 1485 eligible households is too small to create multivariate models and consider, for example, R-indicators or coefficients of variation. Tables 6 a-c displays NL registration rates with and without interviewer for household income in quintiles, household assets in quintiles and educational level of the household reference person. It can be concluded that registration rates are higher for all categories. Especially, for lower income households and lower educations, the interviewers may bring in households that are harder to convince through self-administration.

*Table 6a: NL registration rates with and without interviewer for household income in quintiles Standard errors between brackets.*

|  |  |  |
| --- | --- | --- |
| HH income quintiles | Interviewer | No interviewer |
| 0-20 | 15 (3.5) | 8 (2.3) |
| 20-40 | 16 (3.2) | 15 (3.0) |
| 40-60 | 26 (3.9) | 14 (2.7) |
| 60-80 | 33 (4.3) | 18 (3.0) |
| 80-100 | 41 (4.1) | 26 (3.5) |

*Table 6b: NL registration rates with and without interviewer for household income in quintiles Standard errors between brackets.*

|  |  |  |
| --- | --- | --- |
| HH income quintiles | Interviewer | No interviewer |
| 0-20 | 25 (4.0) | 15 (3.2) |
| 20-40 | 20 (3.2) | 10 (2.3) |
| 40-60 | 30 (4.3) | 20 (3.2) |
| 60-80 | 32 (4.1) | 20 (3.1) |
| 80-100 | 32 (4.3) | 18 (3.1) |

*Table 6c: NL registration rates with and without interviewer for reference person educational level in quintiles Standard errors between brackets.*

|  |  |  |
| --- | --- | --- |
| HH income quintiles | Interviewer | No interviewer |
| Primary | 20 (6.1) | 7 (3.9) |
| Lower secondary | 20 (5.5) | 10 (3.7) |
| Higher secondary | 35 (3.8) | 18 (2.6) |
| Lower tertiary | 41 (5.3) | 33 (4.7) |
| Higher tertiary | 54 (7.2) | 35 (6.6) |
| Unknown | 16 (2.2) | 7 (3.9) |

5.2 INTERVIEWER EVALUATION

After the fieldwork was finished, evaluation sessions were held. Evaluation forms were sent to the interviewers and there was a possibility to talk about theses filled out forms. In NL, a group of 10 interviewers formed a focus group for further investigation of the evaluation forms.

The most important findings from these interviewer evaluations were:

* It is very important to build up a relationship with the respondent. A start questionnaire could be helpful. Interviewers addressed this several times during the focus group. Another tip that was given is to have the interviewer give the conditional incentive at the end of the fieldwork to the respondent in person, rather than sending in via mail.
* In all cases, people at the door were either very enthusiastic to participate and installed the app directly, or they refused.
* Interviewers were given a lot of information, which they could not always explain to the respondents. They often directed the respondent to the extra materials (brochure, website, etc)
* Interviewers did not find people reacting different because this survey was done via an app in the Netherlands. They heard the same arguments as they hear at other surveys when people did not want to participate.
* In Spain, however, especially aged people did not have smartphones so they were not able to participate. Also in Spain, people were not very known with the technology which lead to more refusals.
* Frequently heard reasons for not participating were things as lack of time and lack of interest. The subject of the survey did not really matter.
* Interviewers wanted to know if people needed a motivational call during the fieldwork, for example when the data shows that people find it hard to put their expenses in the app.
* The motivational calls during fieldwork mainly resulted in that respondents experienced no problems at all or that people who intended to take part in the end lead to a refusal.
* September was for fieldwork still a month where several addresses could not be visited due to vacation of respondents.
* Corona was a big issue during the fieldwork in October, with several regulations for the interviewers.
* Not all addresses could be visited because of capacity issues among interviewers.
* None of the interviewers mentioned the corona virus as a reason for non-response.
* Both respondents and interviewers noted that they missed questions about the recurring expenses in the app.
* Interviewers noticed that the decision to participate or not was already made, even before presenting the insights page.
* Interviewers in Spain offered more help for installing the app than the Dutch interviewers. In general, this help was more needed for older people or for people that only had an smartphone that was older than 3 years.

Next to the evaluation session that was held, the paper forms of the motivation calls that were held in the Netherlands were analysed. The main finding here was that either respondents found that everything worked fine and that there were no problems at all, or that respondents in second instance decided not to participate in the survey or decided to stop reporting their purchases. In total, in the Netherlands, 97 calls were held.

Data from the interviewer evaluation reports

* 9 out of the 97 calls were not answered.
* 2 respondents did not want to give a number to the interviewer
* Out of the 97 returned forms, 17 addresses did not want to participate.
* Out of the 97 respondents, 60 installed the app.
* Out of these 60 households, 54 started tracking their expenses.
* Of these 54 respondents, 18 problems were mentioned, 4 where the camera did not work at all, 2 people reported that the login codes did not work, and the other 14 problems all involved problems with specific receipts (large receipts, receipts from abroad, etc.)

These results from the interviewer calls give great insights in the problems that respondents encountered. However, the problems that were reported could not be solved by the interviewers on the phone. This is due to the fact that these problems come from the app itself and not from the respondents’ interaction with it. None of the respondents really needed help with the use of the app. So it can be concluded that the app itself in terms of user interface worked fine and that the issues encountered were only technical.

1. DATA QUALITY
   1. DATA RICHNESS AND DIVERSITY

As mentioned in the first section, data quality of the quantitative data was analysed for 7 indicators:

1. the number of purchases per household (entries);
2. the average amount of each entry
3. the difference between the maximum and minimum amount spent per entry per household (amount variation);
4. the standard deviation of the amount of money spent per entry within a household (SD amount);
5. the difference between the maximum and minimum number of products bought per entry per household (products variation);
6. the standard deviation of the number of products bought per entry within a household (SD products);
7. the expected number of different store types in which a household entered a purchase (store types).

Table 6 shows the means, quartiles, and standard deviations of indicators 1-7, without indicator 2 for the three countries in the 2021 app-based HBS and the 2020 web-based HBS. The app-based samples from Luxembourg and Spain entered more purchases on average than the 2020 web-based sample, but there is no difference in the average number of entered purchases between the 2020 web-based sample and the 2021 Dutch app-based sample. Outcomes of the data quality indicators cannot be compared to the HBS 2020 data because there is no photo option, only manual entries.

Despite removing households with extreme total amounts from the 2020 web-based sample, the maximum amount range is still very high. However, there are no substantial differences in the average amount range among the four samples. This is also true for the average SD amount. This table further shows no differences in the average products range and average SD products between the four samples. Finally, households from the 2020 web-based sample entered more store types on average compared with any of the 2021 app-based samples.

*Table 6: Data quality indicators for NL (compared to regular HBS), LU & ES*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Mean | 1st Q | Median | | 3rd Q | Max | SD |
| Entries | NL (2020) | 10.00 [9.88, 10.12] | 6.00 | 9.00 | | 13.00 | 91.00 | 5.88 |
| NL (2021) | 10.54 [9.65, 11.48] | 6.00 | 10.00 | 14.00 | | 42.00 | 7.35 |
| LU (2021) | 11.66 [10.77, 12.61] | 5.00 | 10.00 | | 16.75 | 53.00 | 8.90 |
| ES (2021) | 12.40 [10.82, 14.03] | 5.25 | 10.00 | | 17.00 | 56.00 | 9.70 |
| Amount variation | NL (2020) | 3.91 [3.75, 4.08] | 1.23 | 1.92 | | 3.42 | 301.55 | 9.04 |
| NL (2021) | 3.68 [3.06, 4.30] | 1.46 | 2.38 | | 3.90 | 35.56 | 4.77 |
| LU (2021) | 3.71 [3.17, 4.31] | 1.05 | 1.98 | | 3.79 | 35.28 | 5.18 |
| ES (2021) | 4.38 [3.17, 6.03] | 1.51 | 2.47 | | 4.10 | 76.27 | 8.32 |
| SD amount | NL (2020) | 1.25 [1.20, 1.30] | 0.40 | 0.66 | | 1.13 | 100.30 | 2.70 |
| NL (2021) | 1.17 [1.01, 1.37] | 0.54 | 0.81 | | 1.39 | 13.33 | 1.40 |
| LU (2021) | 1.11 [0.96, 1.27] | 0.37 | 0.66 | | 1.14 | 10.00 | 1.41 |
| ES (2021) | 1.36 [1.03, 1.78] | 0.51 | 0.86 | | 1.40 | 18.37 | 2.16 |
| Products variation | NL (2020) | 3.90 [3.84, 3.96] | 1.89 | 3.21 | | 5.29 | 28.16 | 2.93 |
| NL (2021) | 3.64 [3.19, 4.12] | 0.64 | 2.98 | | 5.32 | 21.29 | 3.57 |
| LU (2021) | 4.07 [3.55, 4.63] | 0.64 | 2.25 | | 6.26 | 29.55 | 4.96 |
| ES (2021) | 3.74 [3.26, 4.23] | 1.90 | 3.26 | | 5.43 | 14.78 | 2.77 |
| SD products | NL (2020) | 1.38 [1.36, 1.40] | 0.67 | 1.13 | | 1.83 | 14.43 | 1.06 |
| NL (2021) | 1.26 [1.11, 1.42] | 0.33 | 1.05 | | 1.81 | 6.65 | 1.15 |
| LU (2021) | 1.32 [1.14, 1.51] | 0.20 | 0.73 | | 1.91 | 14.99 | 1.70 |
| ES (2021) | 1.30 [1.13, 1.47] | 0.73 | 1.15 | | 1.68 | 4.61 | 0.95 |
| Store types | NL (2020) | 4.49 [4.45, 4.53] | 3.00 | 4.00 | | 6.00 | 11.00 | 1.91 |
| NL (2021) | 4.13 [3.88, 4.38] | 3.00 | 4.00 | | 5.00 | 9.00 | 1.95 |
| LU (2021) | 4.18 [3.97, 4.39] | 3.00 | 4.00 | | 6.00 | 9.00 | 1.96 |
| ES (2021) | 3.81 [3.50, 4.13] | 2.00 | 4.00 | | 5.00 | 8.00 | 1.87 |

Overall, these findings show that for four out of the six indicators there is no substantial difference between the 2020 web-based sample and any of the 2021 app-based samples. Thus, there are no differences in the variability of the amounts or numbers of products entered per purchase between the web-based and app- based budget surveys. The 2020 web-based sample performs better for one of the indicators (store type), but two of the 2021 app-based samples perform better on another indicator (entries). With these slight differences only, it can be stated that the data collection of the HBS app is of at least the same quality as it was for the original paper-and-pencil approach.

6.2 IN-APP BEHAVIOUR

In order to analyse the respondents’ in-app behaviour and to give an indication for their motivation and involvement, two types of data were analysed, being the purchases that were reported in the app, as well as the paradata of respondents. In this section the types of entries and amounts will be discussed first, followed by the paradata analyses.

The following table shows the average behaviour of the respondents per day, split over the two different conditions for interviewer yes/no. To see if there is a difference in purchase reporting behaviour, the following indicators were analysed:

* The number of entries per day, also split into manual/scanned entries
* The amount of the these entries per household, also split into manual/scanned entries

The following table shows the results for these indicators

*Table 7: The number of entries (manual/scanned) and amounts, per household & per day*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Mean | 1st Q | Median | | 3rd Q | Max | SD |
| Entries per day per household | Interviewer | 1.05 [1.0408 - 1.0532] | 0.9990 | 1.0117 | | 1.1041 | 1.4440 | 0,0092 |
| Letter | 1.05 [1.0449 - 1.0566] | 1.0123 | 1.0588 | 1.0907 | | 1.1299 | 0,0051 |
| Manual entries per household per day | Interviewer | 0.7932 [0.7877 - 0.7987] | 0.7578 | 0.7763 | | 0.8181 | 0.9047 | 0,0055 |
| Letter | 0.8438 [0.8380 - 0.8497] | 0.7941 | 0.8456 | | 0.8860 | 0.9314 | 0,0058 |
| Scanned entries per household per day | Interviewer | 0.2538 [0.2517 - 0.2559] | 0.2374 | 0.2471 | | 0.2685 | 0.2918 | 0,0021 |
| Letter | 0.2069 [0.2055 - 0.2084] | 0.1949 | 0.2083 | | 0.2169 | 0.2279 | 0,0014 |
| Amount of the entries per household | Interviewer | 40.18 [39.82 - 40.55] | 36.99 | 40.97 | | 42.85 | 46.34 | 0,36 |
|  | Letter | 61.62 [57.77 - 65.46] | 44.73 | 48.58 | | 52.00 | 150.76 | 3,85 |
| Avg amount per purchase scanned | Interviewer | 38.05 [37.26 – 38.83] | 33.86 | 35.76 | | 39.11 | 56.73 | 0,79 |
|  | Letter | 44.99 [44.13 – 45.85] | 37.09 | 50.16 | | 50.75 | 55.60 | 0,86 |
| Avg amount per purchase manual | Interviewer | 40.66 [4015 – 41.18] | 36.79 | 39.32 | | 43.10 | 51.70 | 0,51 |
|  | Letter | 65.69 [60.96 – 70.43] | 45.30 | 46.59 | | 54.35 | 175.71 | 4,73 |
|  |  |  |  |  | |  |  |  |

As the results show in table 7, it is seen that the number of entries is more or less the same for both groups. When the purchases are split into manual purchases and scanned ones, it is shown that the people recruited by an interviewer reported more scanned purchases than the letter group. On the other hand, the people that were invited by letter only, show more manual entries. The higher number of scanned entries versus manual entries could be caused by the interviewer itself, that at the door was able to give more information about the receipt scanner part more than logically can be done by only a letter. Not only could the interviewer make the respondent aware of the fat that receipts could be scanned, also the interviewer could show how the scanner works to make it more convenient for the respondent to work with it. The letter-only group had to find this out for themselves.

Another difference that can be noticed between the groups is that the average amount per purchase is higher for the letter group compared to the interviewer group. To see if this effect is caused by the way of entering purchases into the app, the average amount for the manual entries versus the scanned entries was analyzed. Here it is seen that for both the manual entries as well as the scanned purchases the letter group reported higher amounts. So this raises the question if the people that are in the letter group might be composed of different people than the group that was approached by an interviewer. However, because of the small sizes of the responding groups, this can not be analyzed. It can not be concluded if this is the role of the interviewer itself or if this is due to the composition of the groups.

It is hard to give an explanation where these differences could come from. If this were to be in the inaccuracy of the receipt scanning, a difference would be noticed here. The difference between the groups for the manual entries is significant.

The results give reason to argue that the interviewer that was added into the approach strategy does not have a satisfying effect. The response is higher, but it is unclear if this is a very different or more heterogeneous group than the letter group? But if the number of purchases is looked at in further detail, this is the same for both groups. The amount that was reported per purchase is even higher in the letter group. So these results do not point at a certain direction. The clinical relevance of the use of interviewer could also be argued here, since the use of the interviewer did not lead to more purchases reported.

The indicators that were looked at to see if there is a difference in motivation and involvement for app use are time spent in the app over time and the number of pages that were used per day in the app. Both indicators are presented in the table below:

*Table 8: Time spent in the app and pages visited*

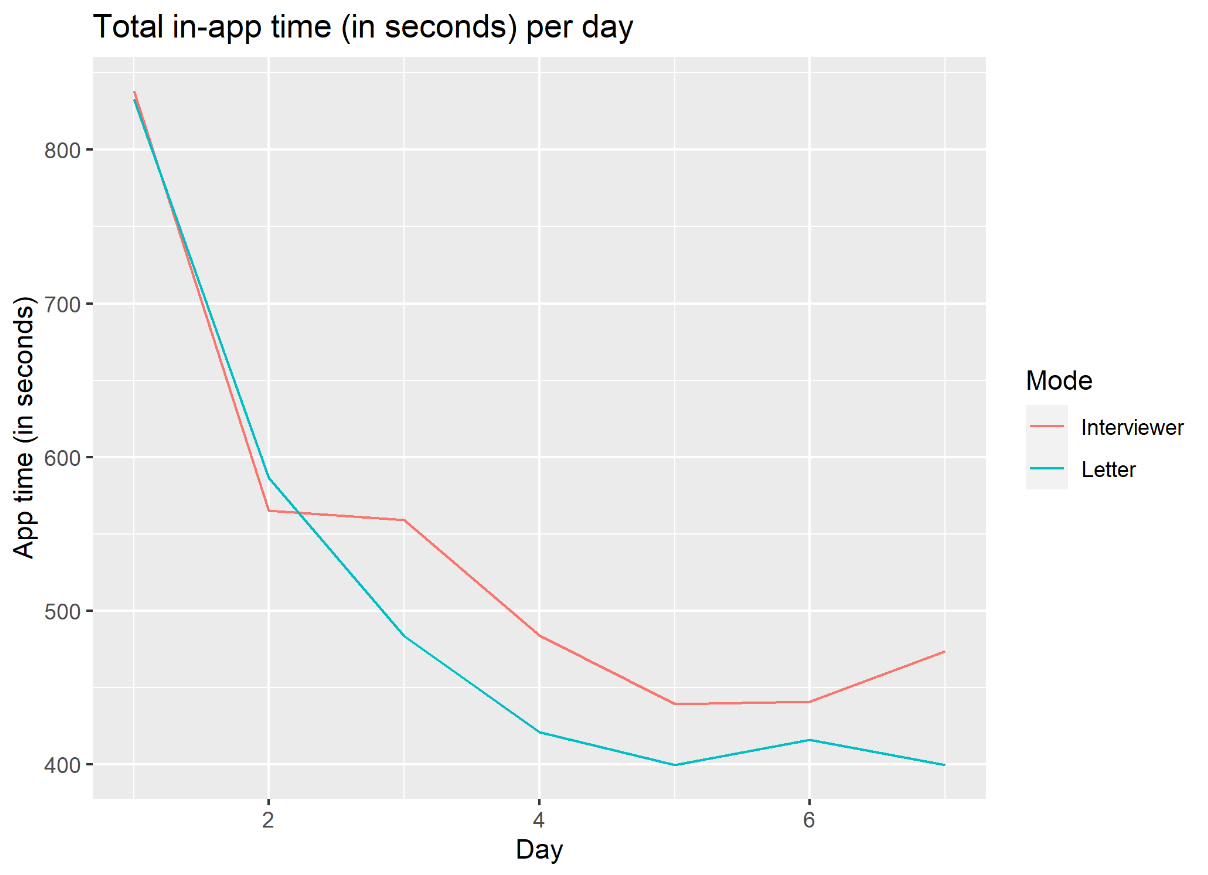
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Mean | 1st Q | Median | | 3rd Q | Max | SD |
| Time spent in the app per day | Interviewer | 542.8 [530.7 - 554.9] | 457.1 | 483.8 | | 562.2 | 838.4 | 12,1 |
| Letter | 505.6 [490.1 - 521.0] | 407.8 | 420.9 | 535.0 | | 833.0 | 15,5 |
| Pages used per day | Interviewer | 20.05 [19.83 – 20.27] | 19.89 | 20.70 | | 21.33 | 22.55 | 0,22 |
| Letter | 19.50 [19.33 – 19.67] | 18.97 | 19.87 | | 19.95 | 22.30 | 0,17 |

One of the most interesting findings is that the table shows that on the paradata the people that were recruited by an interviewer showed more activity in page visits in the app than the group that was recruited via letter only. But it is also noticed that there was no difference in the number of reported purchases. So the lesser time spent for the letter group did not lead to a lower number of purchases. Instead, since the manual entries take more time than the scanned ones, it could be argued that the letter group still invests time to put the purchases into the app correctly. So the next step would be to see if these groups differ in time spent in the app.

Figure 1 shows the time (in seconds) people spent in the app. This was done for both the interviewer group and the letter group to see if there was a difference in in-app time between these groups.

Over time use of the app shows that people invest less time in the app when they are using it for a longer time. This seems logical, since people get used to the use of the app over time. An interesting finding is that the use of the app for the interviewer group is the same for the first few days, and then remains slightly higher for the interviewer group, but the drop shows the same pattern for both groups. So it seems that people in the interviewer group start with the same motivation as the group that received only a letter and that the drop in app use over time is stronger in the letter group. So it could be concluded that the people in the interviewer group stay more motivated over time to invest their time in the app compared to the letter group.

*Figure 1: total in-app time in seconds per day*



But if such a severe drop of app time use is found, the question raises if this also leads to a drop in reported purchases. To further see this, the over-time reporting of purchases was analyzed.

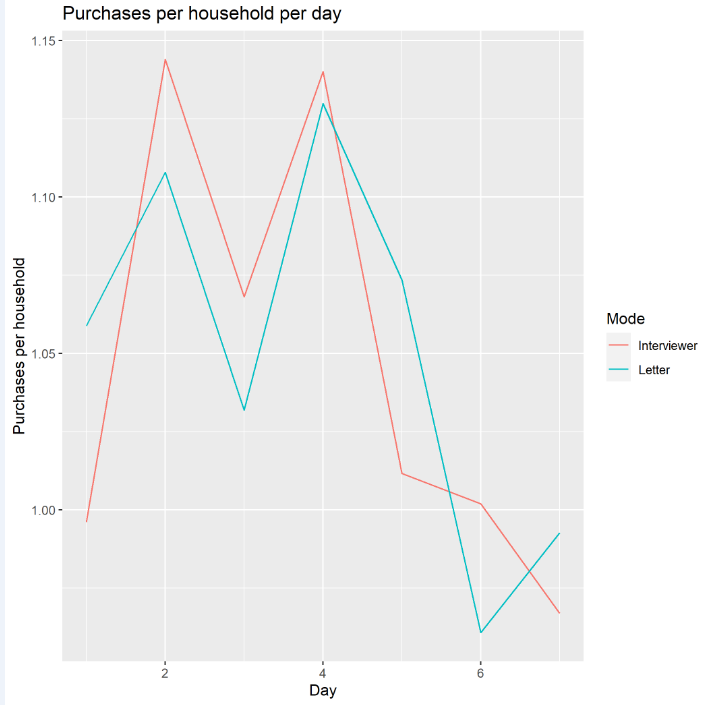
In the figure 2 the over-time use of the app expressed in purchases per household per day can be seen. Figure 2 shows that in the interviewer group people tend to report more purchases in the beginning of the period compared to the letter group. This effect becomes smaller and on some days the households in the letter group reported more purchases per day. So a clear pattern or direction for better purchase-reporting in the interviewer group cannot be concluded.

Over time, a strong declination of the time that the app was used is seen. This can be explained by the fact that people get used to the way the app works and can find their way more and more easy through the app. Also, a decline in the number of purchases reported over time, but the drop of the time spent in the app is bigger than the number of purchases that were reported. It is also clear that on the days that time keeps dropping, the number of purchases increased, i.e. on day 4.

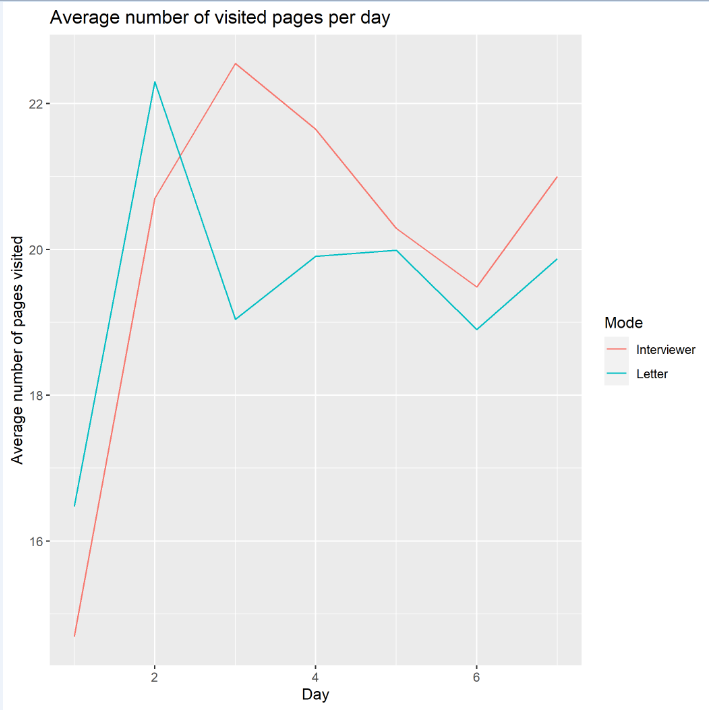
So, from these indicators that were extracted from the paradata it can be concluded that people in the interviewer group seem more motitvated and involved than the group that was not visited by an interviewer. Although this is not seen in the number of reported purchases per day, it is interesting that the interviewer does seem to have an effect on the involvement of respondents. From other surveys, it is known that people that need to be convinced to participate often are less motivated, because they were not intrinsically motivated to participate. because interviewers usually put more effort in convincing the respondents to participate, it is interesting to see that the people in the interviewer group do seem higher involved in using the app.

This higher motivation also effects the number of reported purchases per day, but this effect becomes smaller when time passes.

*Figure 2: number of purchases per household per day*



*Figure 3: the average number of visited pages per day*



1. DISCUSSION AND NEXT STEPS
   1. SUMMARY IMPACT ON REPRESENTATION AND MEASUREMENT

RQ1: what is the impact of interviewer-assistance on registration and completion rates?

The impact of interviewer-assistance on registration and completion rates was someway what it was expected to be. The interviewer assisted approach showed higher response rates compared to the letter group. This was already expected because this is often seen in other surveys. The interviewer at the door can be of great value in convincing people to participate in the survey and to take away their worries and talk them through the app itself. This was clearly seen in the registration rates as well as in the completion rates. Both rates were significant higher in the interviewer assisted group. However, there is no strong evidence that interviewers also recruit different types of households, e.g. older households or lower income households. The number of shared auxiliary variables across the three participating countries was too small for a detailed look.

RQ2: what is the impact of interviewer-assistance on data quality?

In this deliverable, data quality by the effect of the interviewer was looked at through indicators that are predictors of motivation and involvement in the app, such as the number of purchases that were reported, as well as the average amount of these purchases. The number of purchases was the same for both groups, but the amount of the purchases was higher for the letter-only group. Besides that, an interesting effect was found, where people used the scan-function of the app more in the interviewer group when compared to the letter group. So if the in-app data entry option is further developed and perfected, then it is recommendable to have interviewers notice people that this function can be used and that it lowers respondent burden.

It must be stressed that the field test was conducted in unfavourable conditions where part of fieldwork in ES and NL had to be altered or stopped towards the end due to COVID-19. These conditions likely had a negative impact on registration and completion rates.

7.2 ROLE OF THE INTERVIEWER

RQ3: what is the recommended role of interviewers?

So from these results, it can be argued that interviewers do have a positive effect on response rates. It, however, needs further consideration if this would outweigh the costs. The interviewer mode is a very expensive one and does not really affect the composition of the responding groups. Both groups spent a comparable amount of time in the app and also report the same amount of purchases into the app. For respondent burden it could be of great help that the interviewer tells the respondent about the OCR function in the app, so the respondent does not have to put all the expenses in manually.

Next to the quantity or quality of the response, interviewers also can have a helping effect for respondents. They can be the ‘face’ of the NSI, instead of just an somewhat ‘anonymous’ sent letter. Also, when a lot of materials are needed or sent to respondent, it is of great value that there is an interviewer at the door that can give some explanation about what the respondent needs to do as well as the purpose of the research. Even though the interviewer could not really help respondents with technical issues, but for app users it is of great service that they have their personal helpdesk. In this experiment, the role of the interviewer was very small by giving only a letter at the door. Interviewers might have an even greater effect if they also would conduct a start questionnaire.

7.3 NEXT STEPS

All in all, it remains somewhat open whether interviewer-assistance is beneficial. The lack of a broad range of shared auxiliary variables across countries and the potential impact of COVID-19, hamper strong conclusions. Given that higher response rates still are paramount, it is advisable to do so. Main follow-up questions are whether the interviewer-role can be further optimized and whether combinations with other modes are preferable.

To optimize the role of interviewers, a number of steps may be taken: A first step would be to let interviewers get better acquainted with the app. Interviewers did get training and did use the app for a while, but it still was relatively new and exotic to many of them. A second step may be to collect arguments to convince households based on analyses such as in this paper. An important argument may be that the app costs on average two to three minutes per day (depending on the size of the household). Another important argument may be about security and data flows; what is being done with data. Such arguments may best be formulated with the help of interviewers and it is likely that response rates may converge to higher levels with growing experience.

Considering supplemental modes is a natural avenue in many ESS countries. Current HBS implementations often use paper diaries and/or web-based diaries. The data collection strategy may be push-to-app by first offering the app and then switch to one of the other modes. In ESTAT-funded project @HBS2, Stat Slovenia adopted such an approach in the field. See deliverable 3.3 of that project. The sequential mode strategy may improve representation. However, it comes at the cost of comparability and measurement. The data from both paper based diaries and digital diaries, including scanning, need to be combined. It is yet an open research question what are expected mode-specific measurement effects. It is one of the most prominent areas of follow-up research.