Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

Deliverable Report

Deliverable: D3.7
Deliverable Name: Report on keystroke analysis and implications for field work
Deadline: M36
Nature: Report

Responsible: MPG-MEA
Work Package Leader: CITY

Contributing Partners and Editors: Johanna Bristle (MPG-MEA)
Verena Halbherr (GESIS)

Keystroke analysis
and implications for field work

CONTENT

Executive Summary

This deliverable reports on how time measures, recorded in keystroke and time stamp data, can be analysed to inform fieldwork. Keystroke and time stamp analysis is a new field which offers lots of opportunities. Time measures are a valuable tool during *all* phases of the survey lifecycle to inform survey managers - before fieldwork for developing the questionnaire, during fieldwork to check the data quality and post fieldwork for data quality analysis. This deliverable describes analyses conducted during the survey lifecycle of SHARE and after the survey fieldwork in ESS and makes suggestions for further potential analysis.

When working with keystroke and time stamp data it needs to be noted that they are raw data. Hence it is important to evaluate the quality of these data first, that will then be used to analyse data quality of the survey answers. Time for preparation, data cleaning and outlier diagnostic is needed, but worthwhile. The investigation of the distribution of interview length showed that the mode of data collection influences the quality. When interviewers recorded time stamps manually in ESS-PAPI, rounding error occurred. This rounding error is avoided by automatic recording of time measures in ESS-CAPI and SHARE.

Analyses of keystroke data performed during the SHARE pretest showed a slight increase in interview length compared to the previous main fieldwork. To not make the interview longer over waves, durations, variance, and don't know answers of new items were investigated. Questionnaire decisions on inclusion or exclusion of newly introduced items were then made based on these analyses. During fieldwork, length analyses on interviewer level highlighted irregularities in average interview lengths as well as in durations for reading out introduction texts. These pointed survey managers to shortening or skipping behaviour of interviewers. Analyses of time stamp data augmented with respondent characteristics conducted after the end of fieldwork for the ESS showed that interview durations are correlated with education and nationality of respondents. Regarding contact strategies, the analysis of time stamps of day and time of contact attempts displayed that the working population is difficult to contact during daytime and is more likely to be contacted either during the evening or on the weekend.

The analysis of time measures for cross-national flagship surveys like the ESS and SHARE revealed some similarities that seem to go beyond survey-specific peculiarities. In our comparisons on interview length we identified a similar cross-national pattern for both surveys. Therefore, linguistic and country-specific influences need to be taken into account when using time measures for data quality assessment or fieldwork monitoring. Insights from fieldwork analysis can be used to provide guidelines for other surveys which do not yet provide or use keystroke and time stamp data themselves. We recommend analysing time stamps and keystrokes as part of the quality control process of a survey.

1 Introduction

The deliverable "Report on keystroke analysis and implications for field work" of work package 3 on Data Quality (WP3) of the "Data Service Infrastructure for the Social Sciences and Humanities" (DASISH) reports on how keystroke data can be analysed to inform field work. The analyses presented here follow the bi-annual operating cycle of the Survey of Health, Ageing and Retirement in Europe (SHARE) and the European Social Survey (ESS).

In the process of producing survey data much paradata, i.e. data about the process of survey production, are generated. The amount of information on the process of survey production has increased in the last years (Kreuter 2013). Among others this is due to the increasing use and further development of technological means in the context of survey-based data collection, such as computer-assisted personal interviewing (CAPI) techniques, and the implementation of web surveys. Originally the term paradata referred to computer-generated data about the process of survey data collection only (Couper 1998), for example keystroke data or call record information. More recently, a broader concept of paradata has become common, which also includes interviewer observations and information on the interviewers (Couper and Lyberg 2005; Kreuter and Casas-Cordero 2010).

Paradata are key data for analysing data quality in survey production[1]. Most commonly they are used during survey production for monitoring the fieldwork, evaluating interviewer performance, and observing the data production process. If paradata are available on a regular basis during fieldwork they can be used for implementing responsive or adaptive designs[2] to guide data production efficiently and improve data quality (Groves and Heeringa 2006). According to Couper and Singer paradata are used "to evaluate and improve survey instruments but also to understand respondents and how they answer surveys" (Couper and Singer 2013: 57).

Paradata often contain time measures about the survey. In general, this is achieved by setting time stamps during the interview or at least when the interview is started and when it is closed. Time stamps are records of the date and the exact time of day (up to milliseconds). More detailed information can be derived from keystroke data. Keystroke data track all actions on a keyboard and record time stamps for each action taken, e.g. pressing the Return key. The most frequently used information in this deliverable is duration – calculated based on the item-level time stamps in keystroke data for SHARE and recorded time stamps for the ESS. In this deliverable we will focus on the contribution of analysing

[1] For a broader introduction to paradata and how they can be used to inform survey practice, we want to refer to the recently published book Kreuter, F. (2013). Improving Surveys with Paradata: Analytic Uses of Process Information. Hoboken, New Jersey, John Wiley & Sons.

[2] Responsive and adaptive designs are data-driven methods to inform decision-making in fieldwork management. Data from an earlier stage are analysed and implications are implemented in later stages of the fieldwork.

time measures derived from paradata, namely keystroke and time stamp data, to fieldwork in cross-national surveys.

Chapter 2 and 3 provide an introduction to the data types used and describe preparation as well as diagnostic analyses of keystroke and time stamp data. Chapter 4 is the core of this deliverable and contains analyses structured along the survey life-cycle: Keystroke data are used for informing questionnaire development during pretest; for monitoring purposes during fieldwork and for data quality assessment after fieldwork (post-survey). While analyses before and during fieldwork mainly rely on fieldwork experiences in SHARE, post-survey analyses are mainly based on ESS data and make use of own analyses as well as already published survey methodology papers[3]. The analyses of the two surveys intend to complement each other and overlap is limited to two comparisons: Comparisons across the two surveys are conducted for the overall distribution of interview length (Chapter 3.2) and country-specific variations in interview length (Chapter 4.4). A summary of the findings as well as lessons learned can be found in Chapter 5.

2 Data

The main data sources for the analyses in this deliverable are keystroke and time stamp data. In the following we will give a definition of keystroke data and describe the collection, structure and preparation of SHARE keystroke files and ESS time stamp data briefly.

2.1 Keystroke data in SHARE

SHARE uses a CAPI instrument based on Blaise which is centrally programmed by CentERdata and which all participating survey agencies use. While the survey interview is conducted, additional paradata are collected by means of tracking audit trail data, here called keystroke data. This means, every action taken on the keyboard of the laptop is registered and stored by Blaise in a text file. Figure 1 shows the answer options in the CAPI instrument to the question “During the past twelve months, how often did you have contact with your father, either in person, by phone, mail, email or any other electronic means?”. Below the answer options the interviewer can see the section where the CAPI instrument stores the answers (example DN032, which refers to the question text above, highlighted in red). This place on the screen is termed field.

A keystroke extraction of this and some surrounding items is displayed in Figure 2. This is stored as text. In the example highlighted in red, the interviewer selected the answer using the mouse, pressed RETURN (“key:13”), and then the answer which was given is stored. The

[3] Analysis and graphs used with the kind permission of the authors.

text file contains information on the time of entry into a field and exit out of a field, and all actions in between (entering of an answer, editing the answer, opening additional screens like a help file, mouse movements). Time stamps are attached to every action.

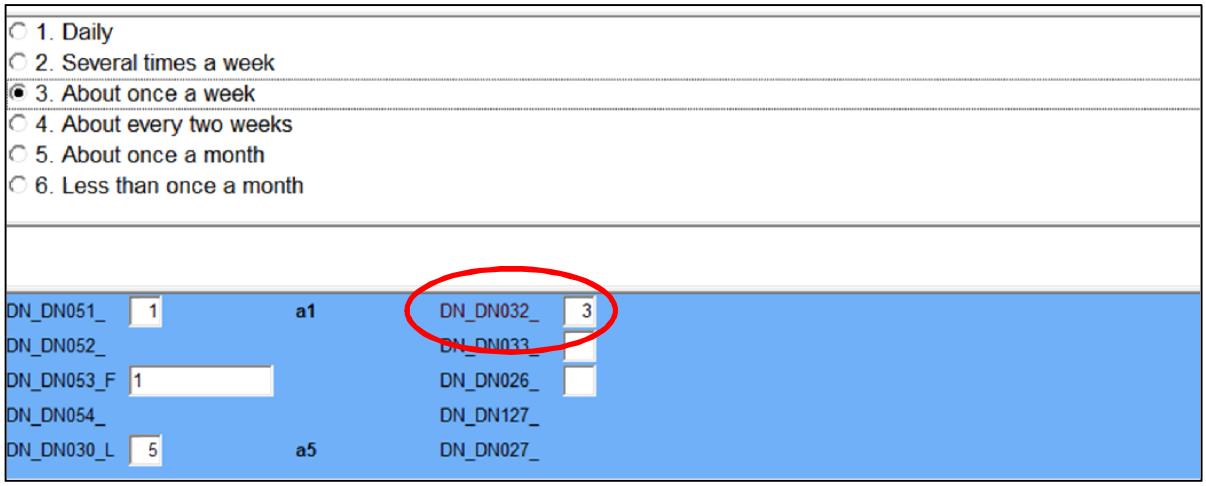

FIGURE 1: SCREENSHOT OF ITEM DN032 IN CAPI INSTRUMENT

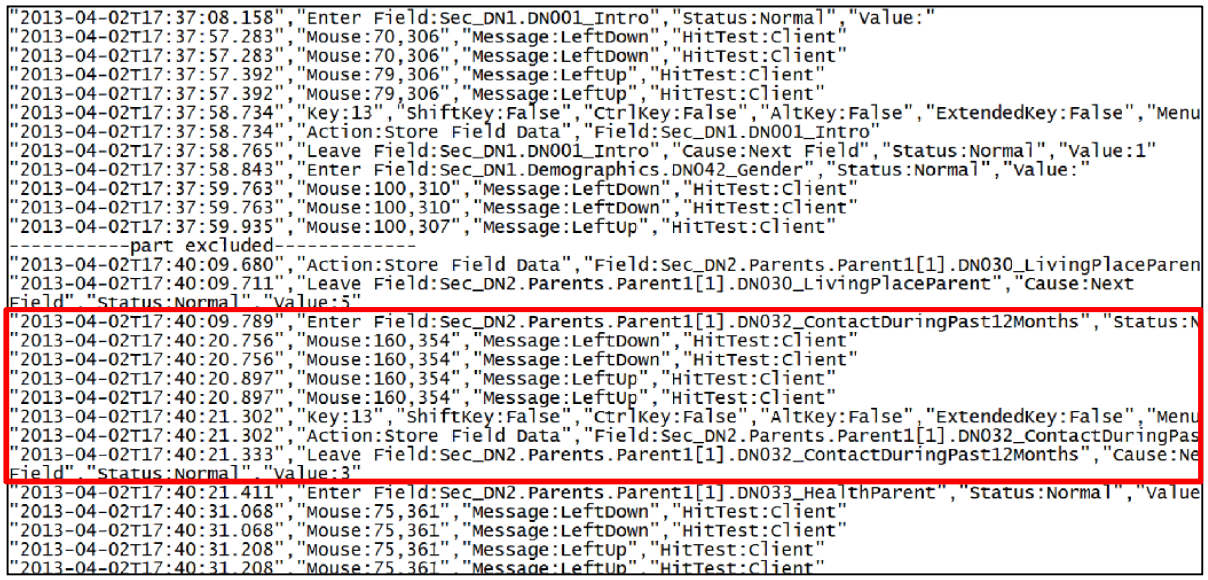

FIGURE 2: SCREENSHOT OF KEYSTROKE EXTRACTION

From these text files, durations on field level are computed by CentERdata and saved as STATA and SPSS files. Besides the time spent on a field, the name of the field (which contains the item number of the questionnaire), the answer of the respondent, if the interview was restarted, the number of times an item was accessed, backed-up, if a remark was set, and the remark itself are recorded. As identifiers to link the keystroke data with survey data, respondent and laptop IDs are extracted. Figure 3 shows a small selection of variables and identifiers (mock IDs used). As an example, during this interview the field which records the answer to the item DN032 was activated for 11 seconds (stored in the variable secfield).

respondent~d	laptop_id	count	field	secfield	remark
DE-000000-01	DE-123	1	Sec_CS.CS011_EffortR	2	0
DE-000000-01	DE-123	1	Sec_DN1.DN001_Intro	50	0
DE-000000-01	DE-123	1	Sec_DN1.Demographics.DN042_Gender	1	0
DE-000000-01	DE-123	1	Sec_DN1.Demographics.DN043_BirthConf	4	0
DE-000000-01	DE-123	1	Sec_DN1.Demographics.DN044_MaritalStatus	4	0
DE-000000-01	DE-123	1	Sec_DN1.Demographics.DN501_NationalitySinceBirth	7	0
DE-000000-01	DE-123	1	Sec_DN1.Demographics.DN504_CountryOfBirthMother	4	0
DE-000000-01	DE-123	1	Sec_DN1.Demographics.DN505_CountryOfBirthFather	2	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[1].DN026_NaturalParentAlive	7	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[1].DN030_LivingPlaceParent	15	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[1].DN032_ContactDuringPast12Months	11	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[1].DN033_HealthParent	10	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[1].DN051_HighestEduParent	11	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[1].DN053_FurtherEduParent[1]	4	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[2].DN026_NaturalParentAlive	6	0
DE-000000-01	DE-123	1	Sec_DN2.Parents.Parent1[2].DN027_AgeOfDeathParent	5	0

FIGURE 3: SCREENSHOT OF SELECTED VARIABLES ON FIELD-LEVEL

The structure of the data is rather complex. Information is stored on field level with one observation in the data per screen that is shown during the interview. A field refers to one item in the CAPI extraction. This eventuates in a hierarchical structure of the data, which in this case are stored in long format: Fields are nested in interviews, which are nested in laptops, which are nested in survey agencies. Furthermore it results in a non-rectangular structure of the data, meaning the number of observations varies over respondents. This is due to the fact that respondents receive different amounts of questions due to interview roles assigned to the respondent[4], the use of preloaded information[5], loops, unfolding brackets[6] and routing.

The volume and structure of the data does not only require lengthy processing and computation times, but also proper aggregation procedures and adequate analytical methods. The aggregation level should always be chosen with regard to the research question and the purpose of the subsequent analysis. For descriptive purposes keystroke data in SHARE are aggregated at the respective level of interest, e.g. item, module, or respondent level. The respondent level file is part of an internal paradata set and stored as a generated variable module (gv_ks). It is cleaned in accordance to the SHARE survey data to match the release data. An overview of the indicators created in the module gv_ks is presented in Table 1 below.

[4] Some information in SHARE is collected on the household level and therefore is only answered by one member of the household .Those interview roles assigned are financial respondent, household respondent and family respondent.

[5] Information from previous waves is loaded into the CAPI instrument when starting the interview with the attempt to ensure panel consistency and time efficiency.

[6] Unfolding brackets are an instrument to reduce item nonresponse. When the respondent does not know or neglects to answer e.g. an income question, a follow-up question is asking if the income is below or above a specific amount. Further follow-up questions of this type will provide a range to give an estimate of the actual amount of income.

TABLE 1: INDICATORS OF THE KEYSTROKE MODULE ON RESPONDENT LEVEL FOR SHARE

Module	Data source	Indicators
gv_ks	Keystrokes (time stamps after each item, tracked in Blaise)	• Length of interview • Number of items asked • Length for each module • Number of items for each module • Last module in case of breakoff • Length of selected introduction items

A well-written exemplary guide on data management of paradata can be found in the aforementioned book on paradata edited by Frauke Kreuter in chapter 4 (Yan and Olson 2013).

2.2 Time stamp data in ESS

The ESS has a decentralized fieldwork structure with a centrally organized core to ensure comparability across countries. This procedure includes (among others) a centrally developed questionnaire, centrally coordinated and verified translation, and specifications which all countries need to fulfil to ensure an input-harmonised survey (European Social Survey 2011). In all of the countries the fieldwork is conducted face-to-face, in some countries in PAPI in others in CAPI. The time measures provided differ between PAPI and CAPI countries. The details on the time used to answer an item obviously cannot be provided in countries using PAPI. Also the software used to conduct the survey differs between the countries. This is a major difference to the software used in SHARE, which is centrally programmed. Therefore, information available in the ESS is more limited than in SHARE.

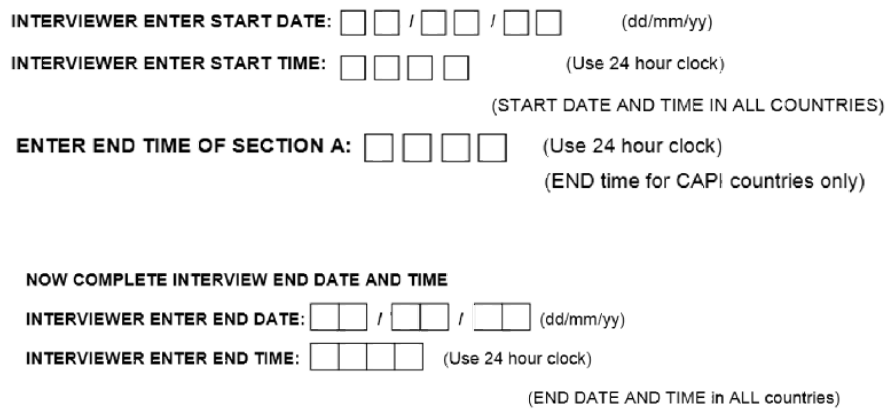

FIGURE 4: PAPI VERSION OF THE ESS 6 QUESTIONNAIRE TO CAPTURE THE BEGINNING AND END OF THE INTERVIEW AND THE END OF MODULES

In the ESS, information on the day and time of the interview and also on the length of the interview is available for all countries. Since Round 6 (2010) the length of each module for countries using CAPI is available additionally. The information is collected automatically for the CAPI countries. In the PAPI countries the information is written down by the interviewer on the paper questionnaire (see Figure 4). The length of the interview is calculated based on start and end time of the interview. The start time and date is recorded just before the first question on the Module A. The end time is recorded after the last question of the main questionnaire (in ESS 5 after G 88, in ESS 6 after F 60).[7] A dataset containing time measures is publicly available and includes the length of the interview and - starting from round 6 – also the length of the modules (see Table 2).

TABLE 2: AVAILABLE INFORMATION ON TIME AND INTERVIEW LENGTH RESPONDENT LEVEL FOR THE ESS

Name dataset	Variable name	Indicators
Main dataset	inwtm inwdds, inwmms, inwyys, inwdde, inwmme, inwyye inwshh, inwsmm inwehh, inwemm	Length of interview (total) Start date of fieldwork (day, month, year) End date of fieldwork (day, month, year) Start of fieldwork (hour, minute) End of fieldwork (hour, minute)
Interview Time dataset (available for ESS 6 only)	inwtm, binwtm cinwtm dinwtm einwtm finwtm inwtm ainwehh, ainwemm binwehh, binwemm cinwehh, cinwemm dinwehh, dinwemm einwehh, einwemm finwehh, finwemm	Length of module A to F End of module a to f (hour, minute)

2.3 Dissemination of paradata

ESS offers access to different kinds of paradata to download along with the main dataset from the website. Since ESS 6 not only the time and the date of the interview, but also the interview length (for all countries) and the length of the modules (for countries using CAPI) is available (see Table 2). These data are available free of charge so every researcher can use

[7] For ESS Round 1 to 6 information on the total length of the interview is available. For ESS Round 6 (2012) the information on the module length are available as well. These are measure before the start of each module for CAPI countries.

this information for own analysis. The dataset can be downloaded for public use from http://www.europeansocialsurvey.org/data/.

The SHARE data are more comprehensive and treated confidentially. Access rights to paradata are decided on a case-by-case basis. “SHARE currently offers the possibility to conduct certain paradata analyses during a visit as a guest researcher, dependent on a prior evaluation of the concrete research project and subject to special conditions of use[8], which are tailored to the intended use of paradata in the context of the respective research project“ (Schmidutz and Bristle 2014: 23).

3 Diagnostic of keystroke and time stamp data

Keystroke data are recorded during the CAPI interview. As any other type of data, keystrokes are potentially prone to measurement and processing error. Data might be missing or the time stamps might be misleading due to a wrong system time of the computer. Also when analysing the data one needs to be careful how to treat outliers and how to properly aggregate the data. When time stamps are recorded by interviewers manually, data might be recorded not as accurate as automatic recording might do. This procedure is prone to other types of measurement error. In the following chapter we inspect keystroke and time stamp data from SHARE and ESS and make cross-survey comparisons where it is suitable.

3.1 Outliers and distribution on item-level

Due to the non-rectangular structure of the SHARE keystroke data, only the activated[9] fields of the questionnaire appear in the data. In other words, non-activated fields due to routing are not part of the dataset. In the keystroke extraction of Germany for wave 5, the raw data file contains 1.85 million observations within 5796 interviews. This results in an average number of about 320 activated fields per interview. Most of them are questions asked to the respondent, but some are also interviewer checks or preloaded information.

The durations stored in keystroke files need outlier diagnostic. Without correcting for invalid data entries, this would result in biased estimates for summary statistics. Especially the mean, which is commonly used for reporting averages, is vulnerable to extreme values. Outliers can occur due to technical errors (e.g. the time stamp was not set correctly) or

[8] Guest researchers are required to fill out and sign a "Statement concerning the use of internal SHARE data including paradata" and an "Obligation of confidentiality" in accordance with national data protection law.

[9] "Activated” means that the CAPI system recorded some activity in this field. This could be the entry of an answer to a question that the interviewer types in or also just “clicking through“. As soon as the field is entered, this activity is recorded.

when the interview process was interrupted (e.g. third person in the room, having coffee, taking a break, lengthy discussion between interviewer and respondent). For trying to get at the "net" interview time, thresholds need to be set. The most common approach is to set the threshold according to the distribution of the item and exclude all items that are outside the 95% confidence interval. A common alternative is to set the threshold at a fixed value.

We adopted a rather conservative approach and only dropped cases which are caused by a technical error or which are so high that the interview seems interrupted. For practical reasons we set thresholds for plausible values at a fixed value and not according to the statistical distribution. This rule satisfies the requirements of fieldwork management. In SHARE wave 5 all items that take more than 1,000 seconds are set to missing except items in modules that are expected to last long (grip strength measurement and asking for record linkage). Here, durations exceeding the threshold are truncated to 1000 seconds (roughly 15 minutes). Furthermore, we set a minimum fixed threshold of 1 second. Durations of zero seconds are set to missing[10]. From the 1.85 million observations, roughly 13 000 have a field duration of 0 seconds and 189 have durations of more than 1,000s (about 15 minutes). This percentage (0.007 %) is recoded to missing values. The distribution is highly skewed to the right with a mode of 3, a median of 8 and a mean of 15 seconds (see Table 3). It is important to consider outliers at every step throughout the analytical process. Not only on the item level, but also on further aggregated levels like module or interview level, outliers occur and need to be investigated and corrected for.

TABLE 3: DISTRIBUTION OF ITEM LENGTH IN SECONDS OVER ALL ITEMS IN SHARE – GERMANY WAVE 5

	Mean	Q05	Q25	Q50	Q75	Q95	N
All fields	15	1	4	8	15	45	1 851 040
Field after exclusion	14	2	4	8	15	46	1 837 657

Time stamp data can be considered as raw data. Outliers need to be analysed for the interview length as well as the length of each module. Also the time, hour, minutes and the date of the interview need further data cleaning. A useful approach for the ESS data is to delete the interviews with a total length fewer than 30 minutes and more than 180 minutes.[11] For the analysis of the module length it seems a good approach to cut the upper and lower 2% of the module length timing. This approach is taken for the analysis in the paper.

[10] Some fields contain valid survey answers which are not necessarily answered during the survey but have been preloaded (e.g. a child's first name or year of birth). The time spent on verifying the preloaded information is stored in the preceding question. Keeping the automatically generated answers would create a bias the summary statistics. For other cases with field durations of 0 seconds inaccuracy in measuring might be an explanation. The field durations are rounded to seconds and therefore could result in zero seconds although the field was activated for 0.3 seconds (as an example). The decision to keep or drop those cases depends on the underlying research question. The zeros are highly interesting for investigating interviewer behaviour such as satisficing or skipping, but they can be misleading for item-level analysis and informing questionnaire design.

[11] This approach is also used by Loosveldt and Beullens (2013) for their analysis on interview length in the ESS.

3.2 Measuring interview length

When using keystroke data we attempt to measure the pure length of the CAPI interview. However, the theoretical concept of interview length is ambiguous and depends on the perspective of the respective actor in a survey. What a respondent means when asking "How long will it take" is conceptually different from what questionnaire developers estimate when designing a questionnaire. Questionnaire development often follows a rule of thumb, which is "four ticks per minute" (Jürges 2005) or in other words, four questions can be asked in one minute. This does of course not contain respondent- and interviewer-specific characteristics which shape the interview length to a large extent. Furthermore, for face-to-face interviews, the respondent might include the overall duration an interviewer spends at the household, which includes some conversation as well as several parts of the interview. This process is depicted in Figure 5.

The conceptual measurement of interview length (Figure 5: Conceptual Measurement of interview length) includes several steps: For SHARE this is a coverscreen to update information on household members and interview eligibility (CV), the CAPI interview(s) and sometimes a paper-pencil questionnaire after the CAPI interview (dropoff).[12] Similar to SHARE, in the ESS the talk at the doorstep, the selection of the respondent, the interview and the supplementary questionnaire/dropoff are part of the respondent and interviewer interaction. For the ESS, time stamps are recorded by the interviewer at the beginning of the interview and at the end of the interview. For the interviewer the length of one interview in the ESS includes additional questionnaires that are filled in without the respondent. Before the start of the interview this is the observational questions on the environment (existence of litter, classification of respondent homes in house or flat …) and after the interview the interviewer questionnaire.

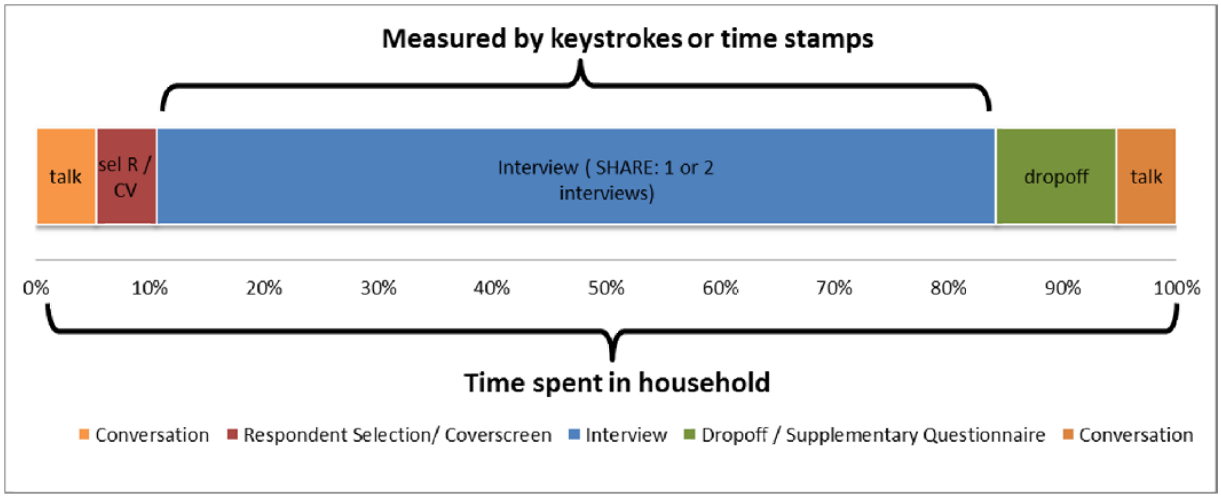

FIGURE 5: CONCEPTUAL MEASUREMENT OF INTERVIEW LENGTH

[12] Only three countries (Austria, Israel and Czech Republic) had a dropoff questionnaire in SHARE wave 5.

In addition, time stamps in SHARE record the length of the coverscreen as well as the overall CAPI length. There are differences in computing the interview length based on keystroke data or on time stamps. With keystroke data it is possible to correct duration measures on item level as described in chapter 2. One example where this procedure is of particular importance is towards the end of the interview when the interviewer section starts (SHARE: IV module; ESS 6: section J[13]). Here interviewers answer a few questions about the interview process without the respondent's participation. Sometimes keystroke data show very long durations at the start of this last module, indicating that the interviewer said goodbye to the respondent and filled out this last module later (e.g. after leaving the respondent's house). Keystrokes therefore provide a more accurate measurement of the actual interview time for survey design purposes.

Table 4 gives an overview of the different time measures on country level for SHARE wave 5. The numbers refer to completed interviews for single households in the panel sample in order to keep the comparison across measurements straightforward and consistent. Outliers were again excluded according to a fixed value. All interviews below 10 minutes and above 200 minutes are excluded from the analysis for the interview length (7 % dropped). The coverscreen length was limited to above 0 and below 20 minutes (1% dropped). The reported average here is the median. Due to missing time stamp data, the number of observations is not consistent with the total number of completed interviews. For comparability reasons, all length analyses in Table 4 are conducted on those cases with valid information on all three dimensions, which is about 98.8 percent[14].

TABLE 4: INTERVIEW LENGTH BY COUNTRY FOR PANEL SINGLE HOUSEHOLDS IN SHARE WAVE 5 (IN MINUTES)

Country	Coverscreen	Keystrokes	Time stamps	Time lag	N
Austria	1.62	61.39	65.30	3.91	1588
Belgium-fr	2.50	77.71	82.26	4.55	896
Belgium-nl	1.97	66.97	71.07	4.10	773
Switzerland	2.05	67.20	71.92	4.73	856
Czech Republic	2.23	66.82	70.45	3.63	1496
Germany	2.10	77.39	82.20	4.81	352
Denmark	1.87	71.08	68.88	-2.20	677
Estonia	2.10	63.76	67.10	3.34	1930
Spain	1.98	58.38	62.78	4.40	879
France	2.27	70.72	74.70	3.98	1792
Israel	1.56	43.77	48.78	5.02	630
Italy	2.13	54.14	58.09	3.95	720
Netherlands	2.28	72.77	75.35	2.58	851
Sweden	2.45	79.18	82.78	3.59	703
Slovenia	1.62	40.40	47.12	6.72	938
Total	**2.05**	**64.77**	**68.37**	**3.60**	**15081**

[13] In the ESS information on the lenght of this selection of the respondent (if applicable), and also the interview questions are not available.

[14] 562 households with completed interviews needed to be dropped due to missing information on one of the length measures. Among the dropped interviews the biggest part was from Germany.

Overall the coverscreen took about 2 minutes per household with some variation across countries. The CAPI interview as measured by the keystrokes took 65 minutes while time stamp data report a length of 68 minutes. This time lag seems adequate taking the above mentioned considerations into account.

Loosveldt and Beullens (2013) investigate time stamps recorded in the ESS. Their results on interview length are presented in Table 5.[15] While in Table 4 above we report median durations for SHARE, the authors report the mean durations for the ESS here. They also investigated outliers and missing data and report the percentage of valid length information in the last column. Valid length varies between 83 % and 100 % per country and is on average at 97.4% (Loosveldt and Beullens 2013, p.70). In addition they report an indicator for interviewer-induced measurement error, which is rounding to the nearest five or ten. This percentage ranges from 18.6 % in Portugal to 85.7 % in Bulgaria and the Czech Republic. The authors note that, in general, more multiple of fives are observed in countries which administered a paper-and-pencil questionnaire.

TABLE 5: INTERVIEW LENGTH BY COUNTRY IN THE ESS ROUND 5 IN MINUTES (LOOSVELDT AND BEULLENS 2013)

	n	Mode	Mean	std. dev.	% multiple of five	Valid length
Belgium	1670	CAPI	60.0	16.5	28.8	98.0
Bulgaria	2427	PAPI	61.6	12.9	85.7	99.8
Switzerland	1435	CAPI	57.3	17.5	19.9	95.3
Czech Republic	2343	PAPI	101.9	27.4	85.7	98.1
Germany	2967	CAPI	76.7	23.7	19.8	97.8
Denmark	1555	CAPI	64.8	18.3	32.7	98.7
Spain	1825	CAPI	69.3	20.8	19.2	96.9
Finland	1841	CAPI	61.6	19.2	19.6	98.0
France	1710	CAPI	68.3	19.1	19.8	99.1
United Kingdom	2336	CAPI	59.5	16.6	20.1	96.3
Hungary	1561	PAPI	68.2	14.2	74.8	100
Israel	1947	PAPI	51.2	15.2	70.0	83.6
Netherlands	1825	CAPI	61.9	18.9	18.7	99.8
Norway	1514	CAPI	67.1	21.0	19.6	97.6
Poland	1731	PAPI	77.2	21.0	60.1	98.8
Portugal	2107	CAPI	55.6	8.9	18.6	98.2
Russia	2595	PAPI	66.4	17.8	62.4	100
Sweden	1484	CAPI	64.4	17.9	21.7	98.9
Slovenia	1348	PAPI	50.6	14.8	75.9	96.1

The distribution of the interview length in the ESS shows the multiple of fives graphically (see Figure 6; calculations and graphical display by Loosveldt and Beullens 2013). The authors interpret this as rounding error. Rounding is a normal way of simplification, but results in inaccurate data. SHARE has a very similar phenomenon when measuring grip strength (Korbmacher and Schroeder 2013; Bristle et al. 2014). Here, interviewers need to read the number on the scale of a dynamometer and then enter it into the CAPI instrument. The distribution of grip strength values shows a very similar pattern with multiple of fives as the interview length of ESS round 5. Figure 7 and Figure 8 show the difference in time

[15] Interviews shorter than 30 minutes and longer than 180 minutes were excluded from the analysis.

measurement for countries using automatic time measure in CAPI software and manual capturing in countries using PAPI in the ESS. In PAPI countries the peaks are due to manual measurement and rounding.

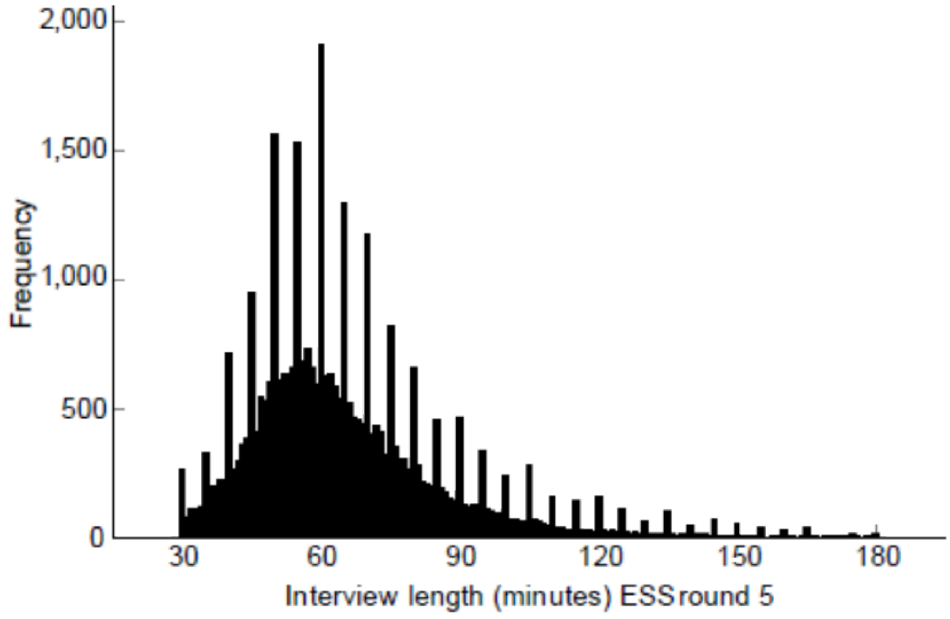

FIGURE 6: DISTRIBUTION OF INTERVIEW LENGTH IN ESS ROUND 5 (LOOSVELDT AND BEULLENS 2013)

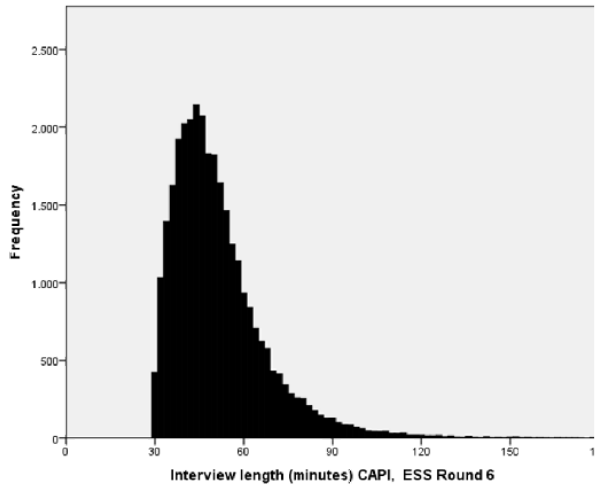

FIGURE 7: INTERVIEW LENGTH ESS ROUND 6 IN IN COUNTRIES USING CAPI

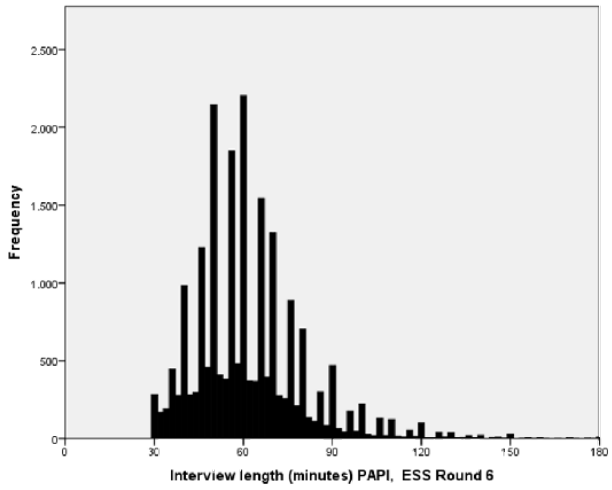

FIGURE 8: INTERVIEW LENGTH IN ESS ROUND 6 IN COUNTRIES USING PAPI

This phenomenon can be observed across items and surveys, but only if humans (here interviewers) are involved in collecting the data. In SHARE, time measures are recorded automatically. The distribution that results from automatic recording can be seen in Figure 9. The distribution is smooth and does not show rounding errors. Consequently, this type of measurement error can be avoided by recording time measures without the involvement of humans.

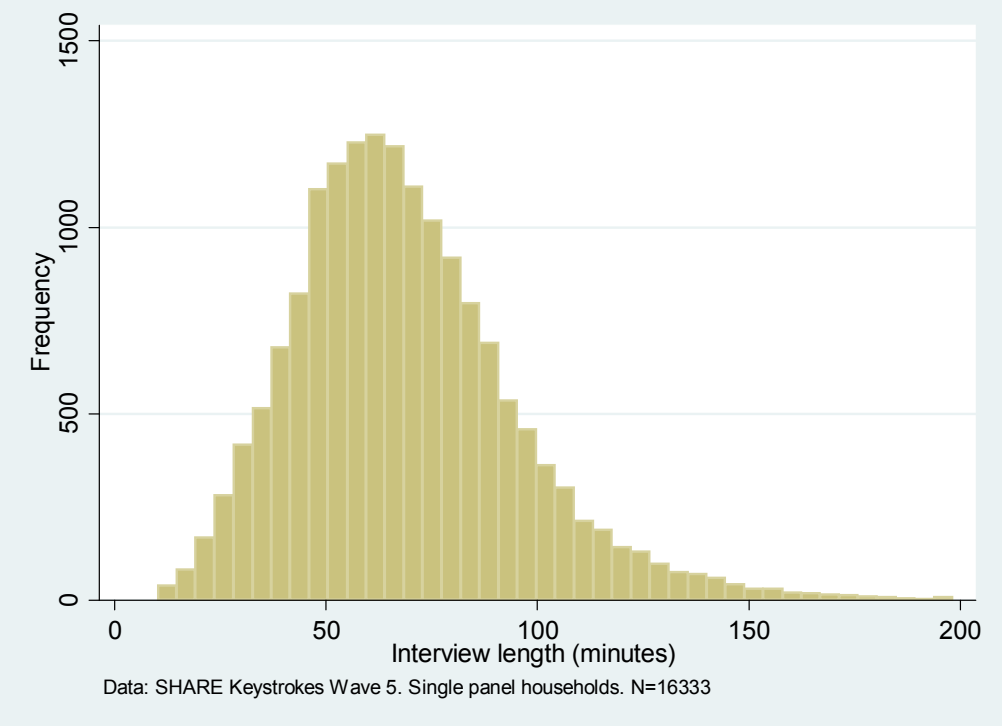

FIGURE 9: DISTRIBUTION OF INTERVIEW LENGTH FOR RESPONDENTS IN SINGLE PANEL HOUSEHOLDS IN SHARE WAVE 5

When comparing the overall distribution of interview length[16] across ESS and SHARE, the similarity of the pattern is striking. Both are only slightly skewed to the right. The ESS Round 5 data interview length has a mean of 68 minutes and a median of 65 minutes (see Loosveldt and Beullens 2013). For SHARE the respective values are at 69 minutes for the mean and at 66 minutes for the median of the SHARE subsample of single households in the panel sample[17]. Similar findings on the distribution of interview length are available for the interview length of the UK Household Longitudinal Study Innovation Panel (Lynn 2013).

The automatic recording of time stamp data avoids measurement error caused by interviewers – here seen in form of rounding. Furthermore, the distribution of interview length is very similar across surveys. The quality of paradata is a less discussed topic in survey research so far. In general, it is important to consider the quality of the original paradata that will be used to analyse and evaluate data quality of the survey data.

[16] The analysis for both surveys excluded extreme values. In the ESS example, interviews below 30 minutes and above 180 minutes are excluded. In the SHARE example, the range displayed is between 20 and 200 minutes.
[17] The comparison of other subsamples of SHARE are displayed in Table 7.

4 Using keystrokes and time stamps in a survey's life-cycle

The analyses presented in this deliverable are of descriptive nature and have been intended to inform the ongoing survey management of SHARE and ESS. The structure of the subchapter highlights the purpose of the respective keystroke analyses and is ordered along a bi-annual survey life-cycle, covering one wave of data collection. Keystroke data are used for informing questionnaire development during pretest (4.1); for monitoring purposes during fieldwork (4.2) and for data quality assessment after fieldwork (post-survey, 4.3). While analyses before and during fieldwork mainly rely on fieldwork experiences in SHARE wave 5, post-survey analyses are mainly based on ESS data and make use of own analyses as well as already published survey methodology papers.

4.1 Pretest: Informing questionnaire development

During pretest the major contribution of paradata is to inform decision-making of the questionnaire development. Questionnaire development in SHARE starts almost two years before fieldwork and changes are evaluated in a pilot and a pretest. Pretest data are analysed in manifold ways and provide the foundation for final decision-making on questionnaire changes. Keystroke analyses can support these decision processes by describing developments over waves or by combining time measures with data quality analyses of item characteristics (e.g. variation on an item).

Interview length

As a panel study, SHARE's main concern in questionnaire development is to balance between keeping the longitudinal dimension on the one hand, and improving or adding measures of substantial interest to the research community on the other hand. Changes in the questionnaire often result in changes in interview length. They are monitored with the overall goal to not make the interview longer over waves. Longer interviews impose a larger burden on the respondents; therefore respondents might be less willing to participate in a survey that takes long. Furthermore, longer interviews are more expensive than shorter ones regarding the payment of the interviewer (Jürges 2005). Over the course of the interview respondent's concentration and motivation might weaken resulting in more satisficing or straight-lining behaviour and reduced data quality (Krosnick 1991). In general shorter item response times and more item nonresponse can be found towards the end of the questionnaire (Galesic and Bosnjak 2009).

In Figure 10 the total interview length of SHARE wave 5 pretest data is presented in comparison to wave 4 main data. Computations are made separately for four subgroups, which represent the major differentiations of SHARE interviews. Here we distinguish

between panel vs. refreshment respondents and between the numbers of interviews conducted within one household (one interview=single, two interviews=couple). It needs to be noted that in a couple interview one interview is most often about the same length as a single interview, while the second interview is much shorter. This is due to routing and the assignment of the roles of household respondent, financial respondent and family respondent. The analysis was conducted based on pretest data and shows that the interview length increased slightly from wave 4 to wave 5. The interview length of the wave 5 main survey was slightly shorter than in the pretest.

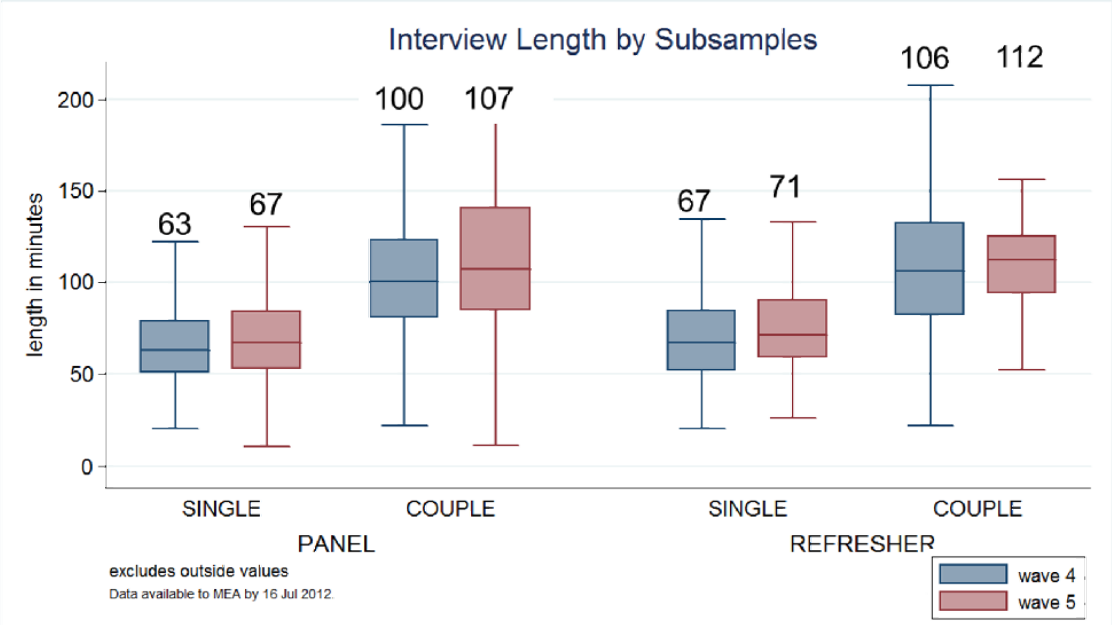

FIGURE 10: INTERVIEW LENGTH SHARE WAVE 4 MAIN VS. WAVE 5 PRETEST

Item length for each item to inform questionnaire decisions

Length analyses are not only carried out on the interview level, but also on item and module level to look at the added overall length for newly introduced and modified items. They are extracted and aggregated on item level (the respective question/item is labelled by the variable name "item3")[18]. An excerpt of this analysis with mean, median, variance (var), standard deviation (sd), minimum (min) and maximum (max) is displayed in Figure 11 and includes pretest data on all countries.

[18] Item level means that durations needed for loops, internal check questions and unfolding brackets within one item of interest are added up per respondent. Afterwards summary statistics are derived (e.g. the mean over all respondents or over all respondents within one country are calculated). At this level, the data was bounded to more than 0 seconds and less than 1,000 seconds (2,000 seconds for the modules grip strength and record linkage).

item3	mean	median	var	sd	min	max
dn001	23,3292	10	2887,18	53,7325	1	1000
dn002	8,10804	5	163,582	12,7899	1	139
dn003	11,1342	7	478,174	21,8672	1	456
dn004	7,27493	3	704,524	26,5429	1	907
dn005	13,5425	9	361,973	19,0256	1	392
dn006	19,1558	11	1207,54	34,7497	1	800
		--- part of output excluded ---				
dn026	12,253	8	763,997	27,6405	1	959
dn027	21,462	12	1426,05	37,7631	1	995
dn028	11,5103	7	499,118	22,3409	1	830
dn029	38,6383	31	984,674	31,3795	1	780
dn030	15,0783	12	317,304	17,813	1	865
dn032	16,2467	13	368,168	19,1877	1	726
dn033	15,1174	11	414,637	20,3626	1	903
dn034	8,33695	5	339,122	18,4153	1	727
dn035	8,42767	6	150,267	12,2584	1	660

FIGURE 11: SUMMARY STATISTICS OF SHARE ITEM LENGTH IN SECONDS (EXCERPT)

One example of how additional data quality analysis is used to inform decision-making is the implementation of social exclusion items in wave 5 (Table 6). Due to routing, the items did not add much time to the overall interview length, but it took on average 4.5 minutes for respondents who received this module. Quality checks on the variance and the percentage of item nonresponse revealed that some proposed items did not perform as expected. Based on the combination of these findings and the length analyses the decision was made to keep a limited number of items measuring social exclusion - those which performed well and in combination stayed within a reasonable amount of added interview time. In the main data collection, 21 items on social exclusion were asked and accounted for 2.5 minutes.

TABLE 6: NEW ITEMS QUALITY ANALYSIS OF WAVE 5 MODULE "SOCIAL EXCLUSION"

Content	Number of items	Variance	Don't Know	Refused answers	Length
Social exclusion	29	Some low	OK	Mostly OK	4 min 30 sec

Data: SHARE wave 5 Pretest. July 2012.

4.2 Fieldwork: Checks for monitoring purposes

During fieldwork the major contribution of paradata is to be able to look at fieldwork progress and interviewer performance on a regular basis and feedback the results to participating survey agencies (and interviewers). Deriving paradata indicators on a regular basis during fieldwork enables data-driven interventions and responsive designs. Given the vast amount of data it is crucial to define beforehand which indicators to look at and which operational consequences might follow.

Fieldwork monitoring in SHARE comprises a broad range of indicators. The data is analysed and reports are distributed on a fortnightly basis with a major focus on cross-country comparison. An overview on fieldwork monitoring in SHARE wave 4 can be found in Malter (2013) and publications on SHARE's methodology (Börsch-Supan and Jürges 2005; Schröder 2011; Malter and Börsch-Supan 2013). In this deliverable we will specifically focus on the contribution of keystroke data to fieldwork monitoring. The structure of this subchapter follows along the analysis level – moving from a broad level (interview length on household level over all countries) to a very detailed level of investigation (item length per interviewer).

Interview length across subgroups and across countries

To obtain high quality data, standardized interviewing is essential in cross-national quantitative survey work, so respondents answer questions under the exact same conditions across interviewers and across countries (Groves et al. 2004). Looking at interview length is a valuable tool to point survey managers to cases which need further investigation, e.g. data quality checks on the actual survey data. Analyses are made separately for the aforementioned subgroups panel vs. refreshment and single vs. couple interview. The overall median for single interviews in wave 5 is 66 minutes in the panel and 77 minutes in the refreshment sample. Mean values are slightly higher (see Table 7).

TABLE 7: INTERVIEW LENGTH SHARE WAVE 5 BY SUBGROUPS

Level	Sample	Median	Mean	Std. Dev.	N
Single HH	Refreshment	77.17	80.27	30.20	4 669
Single HH	Panel	65.90	69.38	26.57	13 987
Couple HH	Refreshment	120.64	123.56	42.41	4 011
Couple HH	Panel	104.60	108.03	37.30	11 913

Data available to Mea by 20 August 2013.

In a cross-national survey, an obvious level of comparison is the cross-country perspective. On the one hand, length variation is due to language and cultural differences which is

legitimate variation. On the other hand, it might be partly due to differential survey management strategies, training or interviewer behaviour, survey climate – causes for variation which are attempted to be minimized in ex-ante harmonised, international data collection. Interviewers play a very important role in face-to-face surveys. For example, in Slovenia and Israel a low total number of interviewers conducted the SHARE study. This means for those few interviewers the interviewer workload was high. In this case it is especially important that the interviewers are trained to conduct the interviews properly and in a standardised way. Otherwise a single interviewer's non-standardised behaviour might affect a relatively large share of the sample.

The large variation of the interview length between countries is consistent across the panel dimension of SHARE. In wave 1 analyses, "the shortest interviews were made in Austria, Spain, and Italy (…). The longest interviews were conducted in Denmark and Sweden" (Jürges 2005: 83). This pattern is repeated in wave 5[19] (see Figure 12).

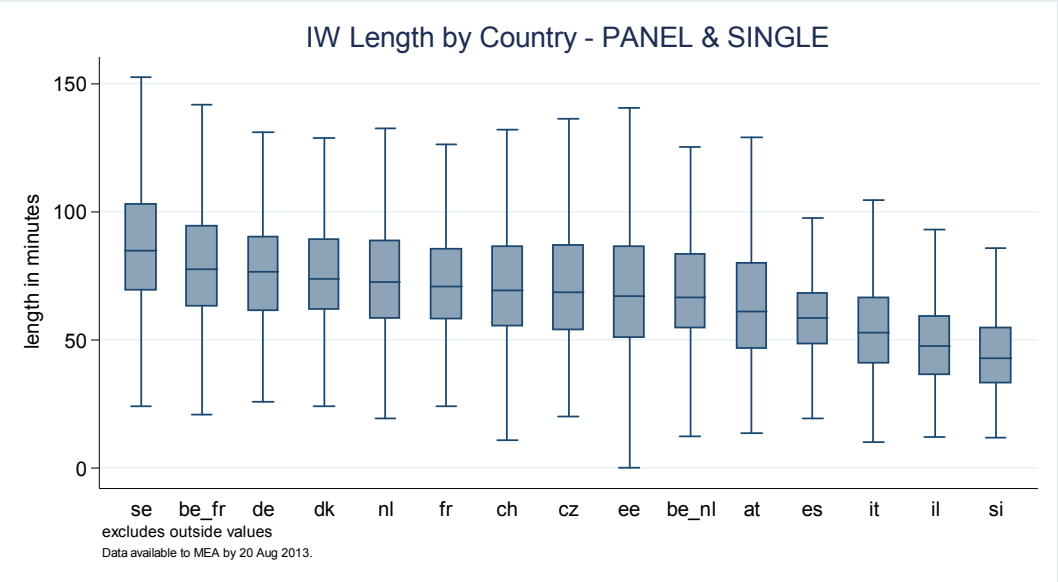

FIGURE 12: INTERVIEW LENGTH BY COUNTRY FOR SUBGROUP PANEL & SINGLE, SHARE WAVE 5

Reading out introductions on interviewer-level (Germany)

Proper reading out of introduction texts is a feature of good interviewer behaviour, which is one indicator for standardisation of data collection (Figure 13). We therefore compared time spent on reading out "long" introduction texts with normative standards (red line, Figure 13). Proper reading would result in a boxplot which is rather high and short. This means it would be centred on a rather high median (close to the normative standard) and

[19] The only countries which show shorter or longer durations in wave 5 than the countries mentioned did not participate in wave 1 (Luxemburg, Slovenia, Israel) or were not part of the wave 1 keystroke analysis (Belgium).

shows only some variation and rather short whiskers (an example is encircled with a dashed line). Ideally there would be little variation for introduction texts. In our case, the item DN001 was differently phrased for panel and refreshment cases and therefore some variation is introduced by design. The question text is as follows:

> *"Let me just repeat that this interview is confidential. Your answers will be used only for research purposes. If we should come to any question you don't want to answer, just let me know and I will go on to the next question. Now I would like to begin by asking some questions about your background."*[20]

For panel respondents the consent to using preloaded information is asked for in addition:

> *"During our previous interview we asked you about your life. To shorten our interview today, I would like to refer to your previous answers instead of asking everything again. Would that be ok?"*

Interviewers who show very little variation around a very short duration do not seem to follow standardised interviewing (encircled with a solid line). Consistently short durations are an indication for skipping behaviour. Their behaviour is monitored further with other indicators on data quality such as item nonresponse.

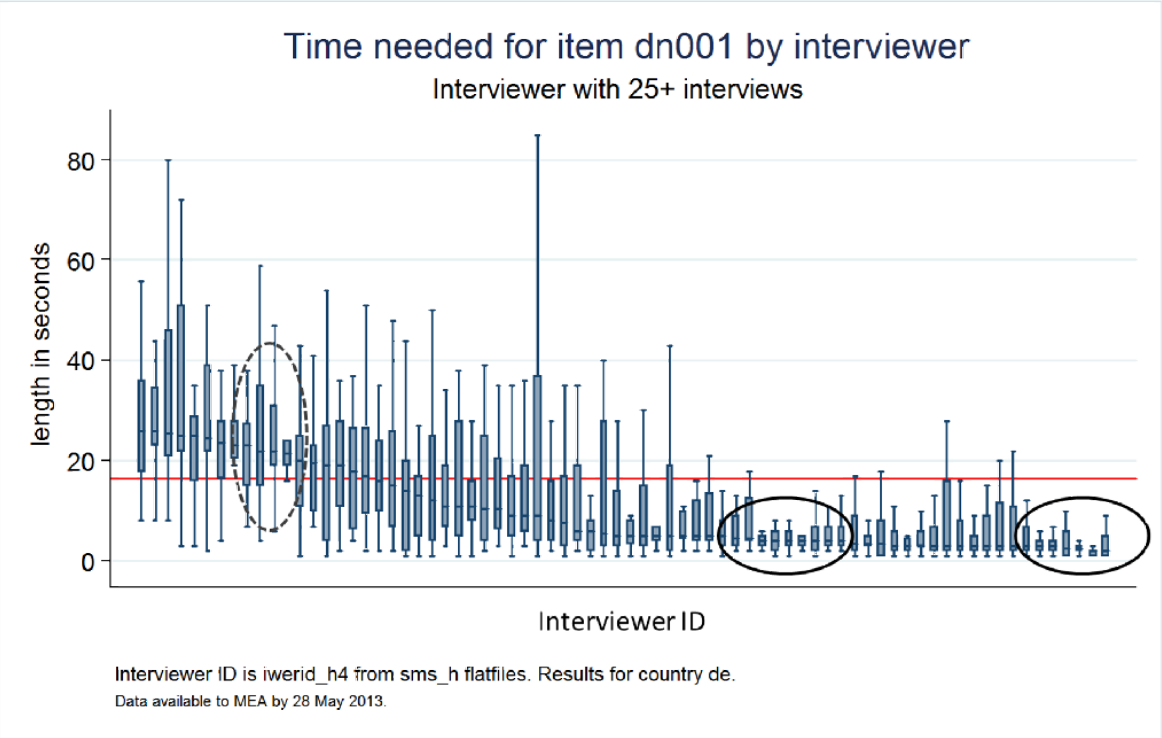

FIGURE 13: TIME TO READ OUT ITEM DN001 FOR INTERVIEWERS WITH 25+ INTERVIEWS IN SHARE GERMANY

[20] The questionnaire can be found online on http://www.share-project.org/fileadmin/pdf_questionnaire_wave_5/SHARE_paperversion_5_4_10_en_GB.pdf.

4.3 Post-survey Analysis: Data quality assessment

In the following chapter we will describe the post-survey use of keystroke and time stamp data with the main focus on data quality assessment. After fieldwork the major contribution of paradata is to add additional information to assess data quality. Of course data quality assessment is not limited to the time after fieldwork. Most analysis can also be done before or during fieldwork, if data are available. In the ESS time stamp data are only available after fieldwork which is why the analysis is mainly limited to the time after the survey. Due to the centralised software system data from SHARE, data are available also for the pretest and during fieldwork. The results from the post-survey checks can be used for informed questionnaire development in the next round or wave of the study and can provide additional insights into the fieldwork process and also into survey quality.

Analyses for subgroups of respondents in different surveys or in different countries can help to assess data quality. Analyses of the total interview length, the length of modules and items can give information on difficulties or differences in language and survey climate. Besides the legitimate variation due to language and cultural differences, other differences might be due to survey management strategies, survey climate and interviewer behaviour (see also chapter 4.2 on other reasons for variation in interview length). Variation could for example also be due to different training of interviewers regarding the handling of «Don't know» answers or reading introductory senctences, which leads to non-standaradized interviewing.

4.3.1 Questionnaire and modules

Cross-national analysis

Information about the length of the survey or the modules can for example provide additional insights on the difficulty of modules in certain countries or for subgroups of respondents. Asking many questions for clarification leads to a longer interview time (see also Loosveldt and Beullens 2013). So the length of modules or the number of questions asked per minute (interviewer pace), could indicate if questions are understood correctly or if interviewers have to give additional help and explanation on certain items. This could guide researchers to further investigate the modules in additional pretesting, e.g. with cognitive pretesting to reveal more of the subject's interpretation of certain items.

In cross-national surveys variation in module length, relative to the overall interview length, might indicate how difficult a module is perceived in the different countries (either for the respondents or the interviewer). As we can see in Figure 14 the overall length of the questionnaire varies between countries. The mean interview length in ESS Round 6 is 55 minutes (for module A to module F). The average interview length in Hungary is the shortest

with 41 minutes and in Cyprus it is the longest with 68 minutes. When looking at the length of module A "social trust and TV watching" across countries with similar overall lengths, we see that Finland and Spain have shorter module A durations than for example in Portugal or Albania, although all four countries have a similar overall interview length. The same pattern holds for the comparison of module B "Politics" between Germany and Cyprus. Although Cyprus has a longer overall interview length, the length for module B is shorter than for Germany. This could also be an indicator for different complexities of the respective module topics in different countries. Besides the length of the wording or the speed of the language other factors like the survey climate could be part of the explanation of variance between countries. Not only do respondents influence the length of the interview, but also the interviewer itself. The length of the interview and the modules can guide researchers to further investigate the differences between countries and interviewers.

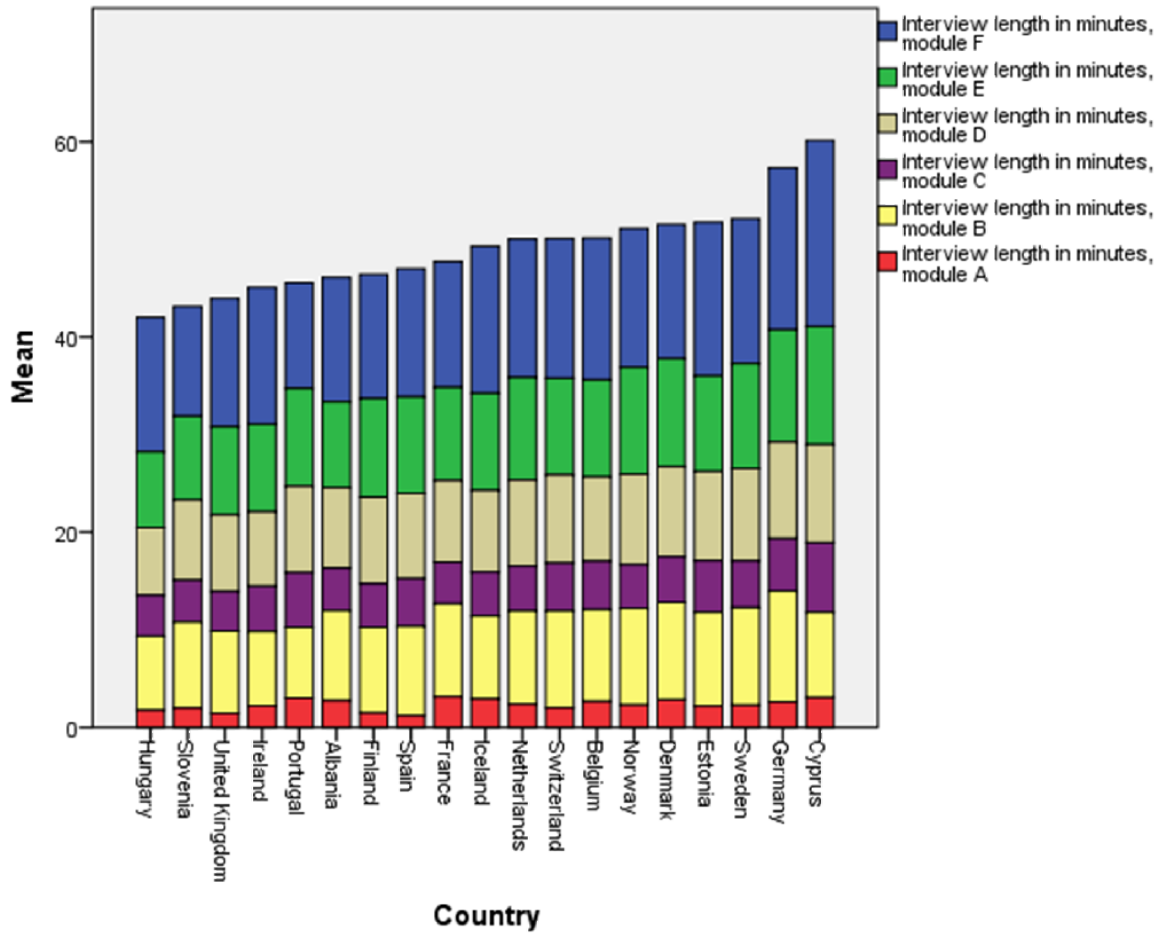

FIGURE 14: INTERVIEW LENGTH AND MODULE LENGTH ESS ROUND 6

Length analysis for different respondent characteristics

Keystroke data provide manifold ways to look at data quality, especially when augmented with further information. For further analysis we look at the interview length for different subgroups of respondents. In a comparison in Germany between the different levels of education we see that people with higher education need slightly more time to answer the questionnaire (Figure 15). On average people with high education needed 64.4 minutes, people with low education and mid education need 2 minutes less (mean low education: 62.6 minutes, mean mid education 62.4 minutes; mean length of interview in Germany: 63.1 minutes).

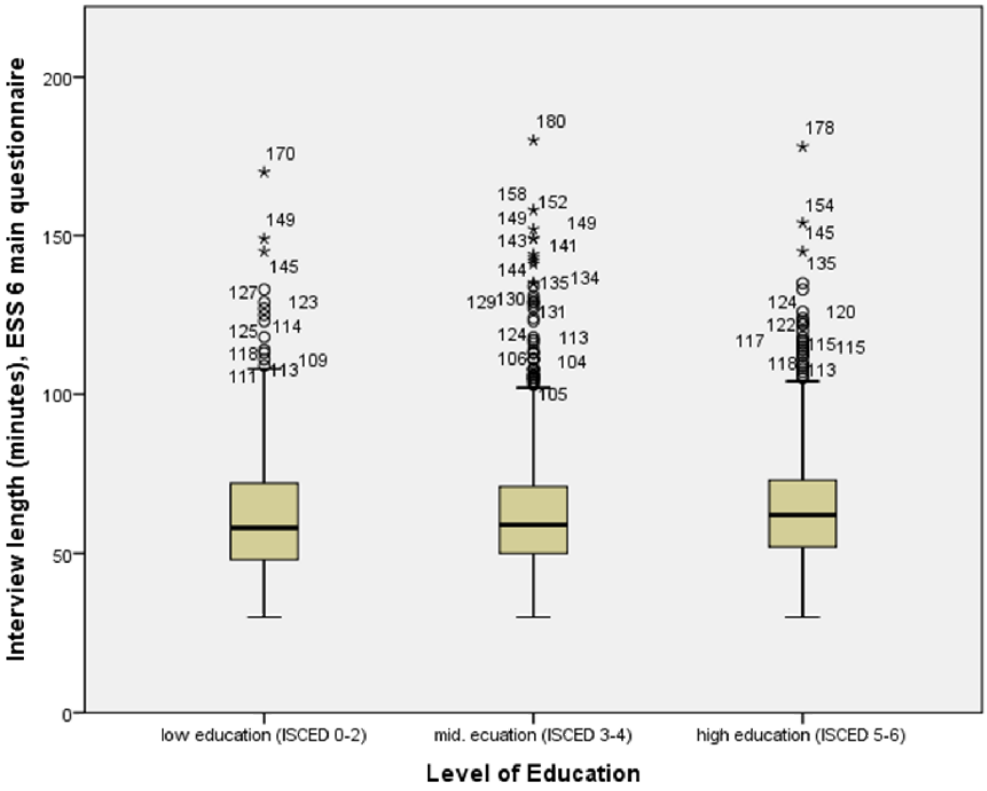

FIGURE 15: LEVEL OF EDUCATION AND LENGTH OF INTERVIEW, ESS ROUND 6 GERMANY

In a second step the length of modules is analysed in combination with survey variables at the example of Germany. No significant differences can be seen for education levels regarding the length of module B on politics, module C (subjective wellbeing), or D (personal and social wellbeing). For the module A on social trust and TV watching, module E on understanding democracy (Figure 16) and F on the socio-demographic profile the interview duration varies significantly over the different levels of education[21]. We have to add that, on the one hand, the length of the module might be an indicator of the difficulty of the questions. On the other hand, the modules include filter questions, which means that not all of the respondents may need to answer all questions. For the socio-demographic module F

[21] Analysis was performed using ANOVA. Results for module A: F(2, 2695) = 3560, p = .029; module E F(2, 2898) = 5722, p = .003; module F: F(2, 2768) = 48065, p = .000.

persons living in a bigger household, and persons with an occupation need to answer more questions than unemployed persons living in one-person households. Since having an occupation is related to the level of education, the number of questions and therefore the length of the module is related to the level of education as well. Therefore this analysis just gives a rough indicator for the difficulties of the modules because it does not control for the number of questions.

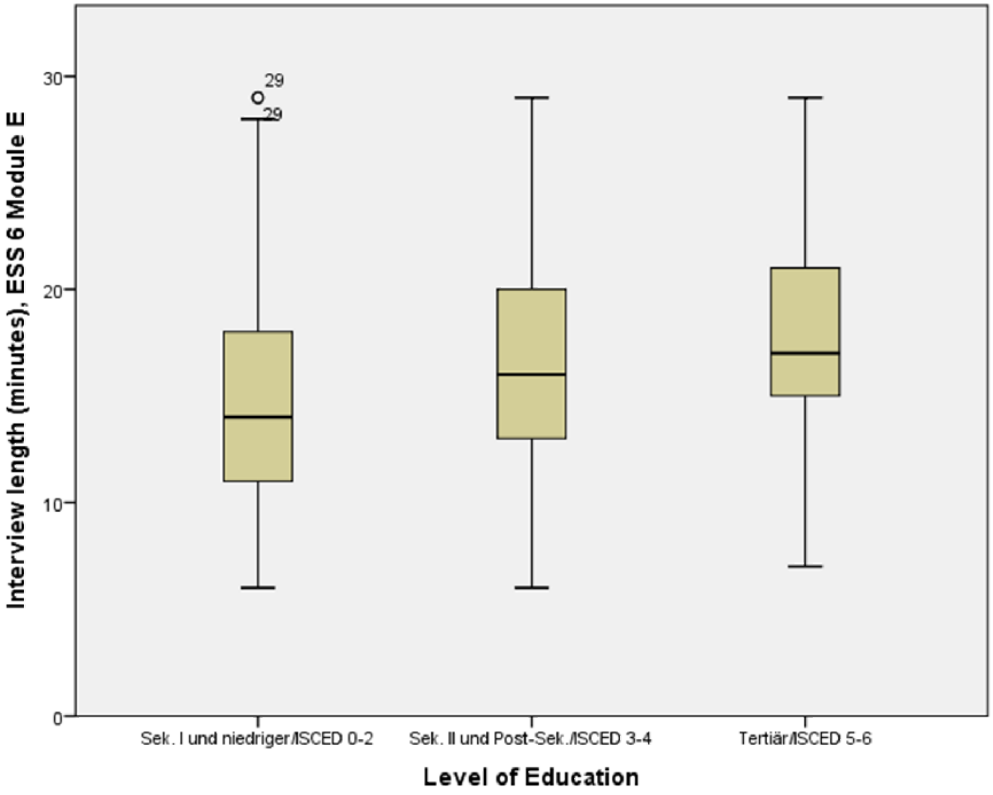

FIGURE 16: LEVEL OF EDUCATION AND LENGTH OF MODULE E (UNDERSTANDING DEMOCRACY), ESS ROUND 6 GERMANY

The length of the interview also differs significantly between nationals and non-nationals in Germany (Figure 17). The mean interview length for a person with German nationality is 62.7 minutes; non-Germans have an interview length of 69.6 minutes. Across all countries the mean interview length is 55.0 minutes for nationals of the country; for non-nationals it is 57.6 minutes. The mean interview length is significantly different between nationals and non-nationals ($F(1, 2919) = 17755$, $p=.000$). For Module B, C, D the non-nationals need significantly more time to finish the module. Overall, the differences in respondent characteristics explain some variation in interview and module durations.

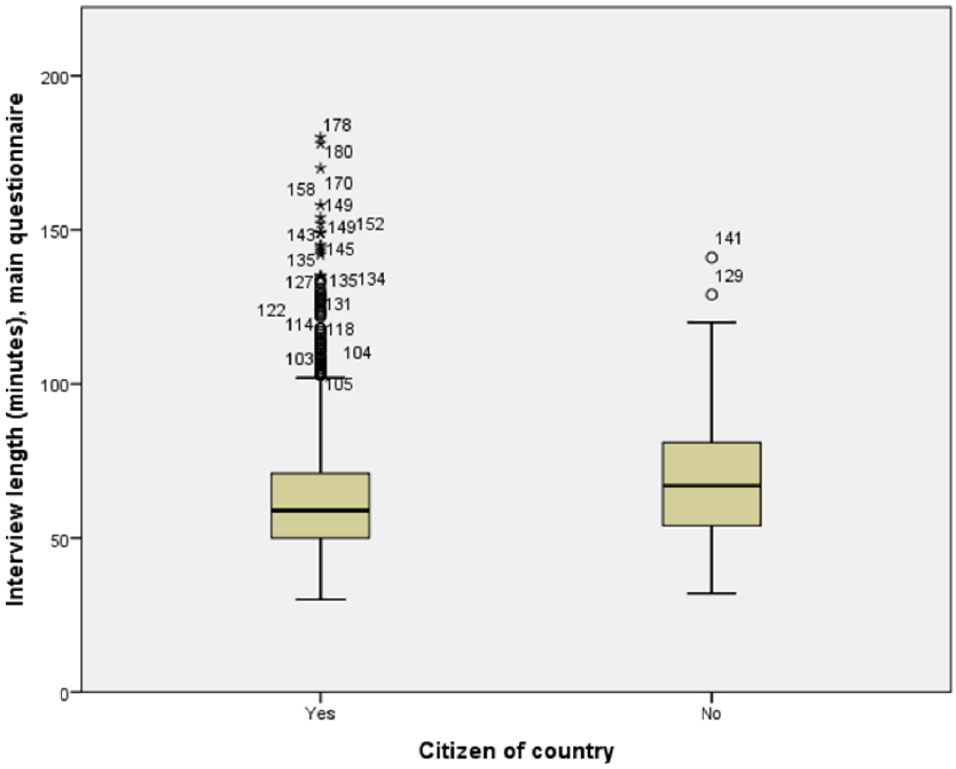

FIGURE 17: LENGTH OF INTERVIEW FOR CITIZEN OF COUNTRY, ESS ROUND 6 GERMANY

Further potential analysis

Loosveldt and Beullens (2013) showed that, in addition to respondent and country variation, the interviewer accounts for about one third of the variance in interview length. Further analysis on the interview length by interviewer characteristics might provide additional insights.

The effects of the interviewer on interview duration analysed by Loosveldt and Beullens (2013) can be further developed. The investigation of interviewer characteristics can add an additional layer to the research on the respondents' effects on the interview duration. Response patterns like satisficing, speeding and shortcutting are mainly analysed from the respondents' perspective (see Krosnick (1991) and Tourangeau et al. (2000)). More analysis on the role of interviewers on respondents' answers and response styles in face-to-face interviews can be conducted.[22]

Additional analysis of short interviews or short duration of module can be conducted to investigate on the data quality of items. One can assume that in short interviews the item nonresponse rate is high, skipping of questions, satisficing or speeding through the questionnaire occur more often.

[22] For analysis on the role of the interviewer on acquiescence see Olson, K. and I. Bilgen (2011). "The role of interviewer experience on acquiescence." Public Opinion Quarterly **75**(1): 99-114.

4.3.2 Plausibility checks and interviewer abnormalities

Keystrokes and time stamps allow checks for plausibility of the time and date of the interview. Analyses of interview length on the interviewer level can be helpful to detect interviewer abnormalities. Also, the data quality of the interviews with very short or long duration might be checked. So the analysis of time stamps and keystroke can add another aspect to plausibility checks and interviewer effects.

Interview length on interviewer-level

In the cross-country comparison within SHARE, Germany has rather long interviews on average with reasonable variation within the sample. Besides questionnaire routing, interviewer behaviour might drive the within-sample variation. For investigating this further, the German country team looked at interview length on interviewer level. In Figure 18, the mean interview durations per interviewer are plotted against the total number of conducted interviews per interviewer. The red, dashed line marks the average interview length for Germany (mean). Overall there is reasonable variation in the amount of conducted interviews as well as in the length. Interviewers who have a high number of interviews and at the same time a low mean duration are encircled in red and might need further investigation. Reasons for short interview duration might be that those interviewers have learned to speed through the questionnaire, skip filter questions or rephrase questions in order to shortcut the interviewing process. Obviously these are non-standardised interviewing techniques and analysing these cases can help to avoid this.

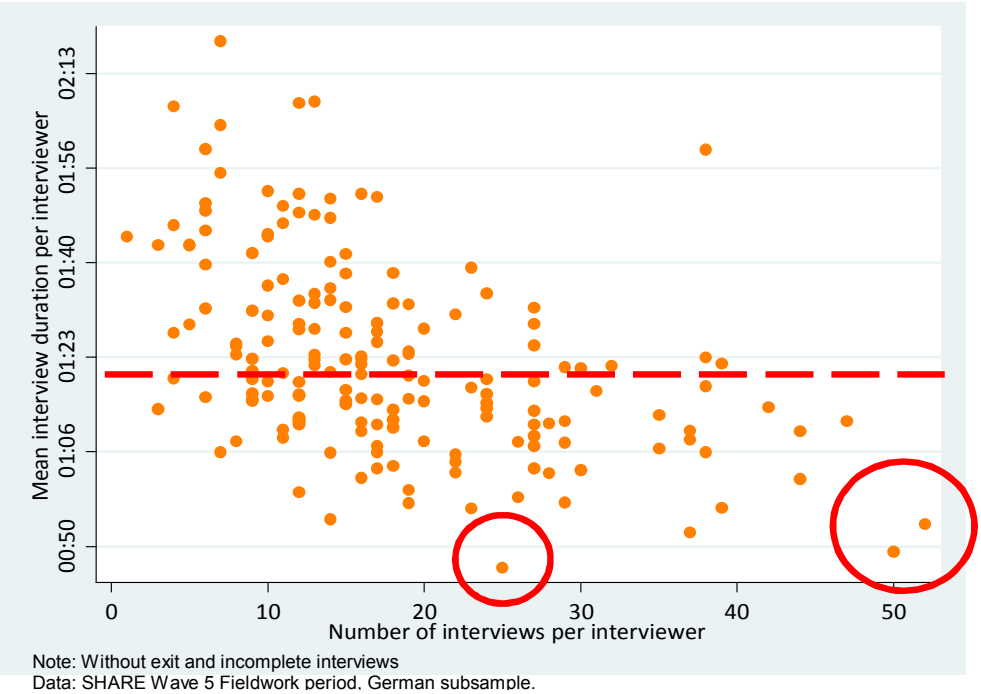

FIGURE 18: INTERVIEW LENGTH PER INTERVIEWER IN GERMANY SHARE WAVE 5

Further data quality checks were performed on the subsamples of those interviewers who showed conspicuous mean interview durations. Firstly, open-ended questions were checked on non-meaningful data entry (e.g. “x” or “abc” instead of children’s’ names or occupation descriptions). Secondly, skipping behaviour was investigated by checking if physical measurements were conducted or skipped. Thirdly, proper reading aloud of introduction texts was analysed using keystroke data again (for details see Figure 13 in Chapter 4.2). The combination of those data quality checks provided the survey managers with an empirical basis for interventions with the interviewers concerned.

Further potential analysis:

- Length of fieldwork period. Example: Are interviews conducted in the relevant fieldwork period? Are interviews conducted before or after the agreed period?
- Time and date of interview. Example: Are all interviews conducted the same day, or even all with only 5 minutes break between the interviews? Interviews conducted at the same time, or with a very short time span between interviews point to suspicious cases. Time stamp analysis allows checking for interviewer fraud (impossible close interviews or interviews at the same time).
- Also analysis of time stamps or keystroke data allows monitoring processes (as described in Chapter This includes checks on interviewer performance, the .(4.2 interviewers’ workload within a certain time range, the interview length within an interview aswell as over the fieldwork period.
- With increasing number of completed interviews per interviewer the length of the interview decreases (Loosveldt and Beullens 2013) (see also Figure 18 in the following subchapter). This could point to shortcutting or skipping of filter questions, speeding and not reading the questions properly.

4.3.3 Insights into fieldwork activities

Analysing the time and date of the interview provides deeper insight into fieldwork activities and into understanding the process of data collection. The collection and analysis on the weekday and the time of day when the interview was conducted can be used in future rounds for planning the interview schedule. Koch and Blohm (2006) analysed day and time of the interview and showed that there are differences between countries (Figure 19). In Finland, Sweden, the Netherlands and France more than 60% of the interviews are conducted during weekday daytime (wkday_dt), while in Israel and Portugal less than 30% are conducted during weekday daytime. A relatively large share of evening calls (wkday_ev) is documented for Israel, Norway and Luxembourg. Of course factors like the employment situation in a country as well as of the particular interviewers might contribute to these differences. While in Finland the interviewers are usually full-time employed by Statistics Finland, in Portugal and Israel the interviewers are part-time freelancers.

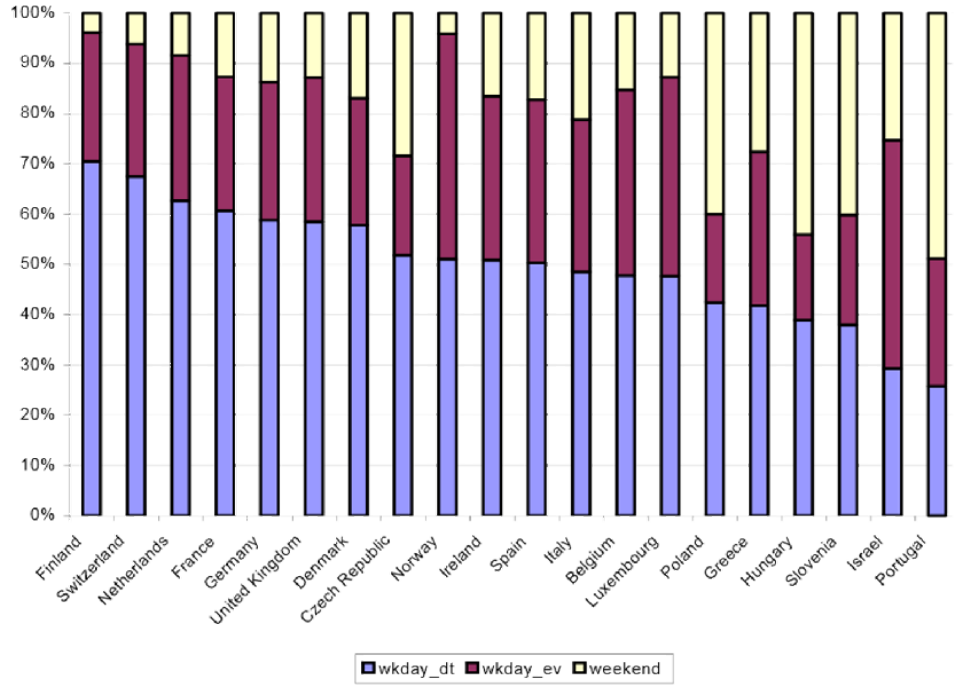

FIGURE 19: DAY AND TIME OF INTERVIEWS IN ESS ROUND 1 (IN PERCENT) (SOURCE: KOCH AND BLOHM 2006: FIELDWORK DETAILS IN THE EUROPEAN SOCIAL SURVEY (P. 32))

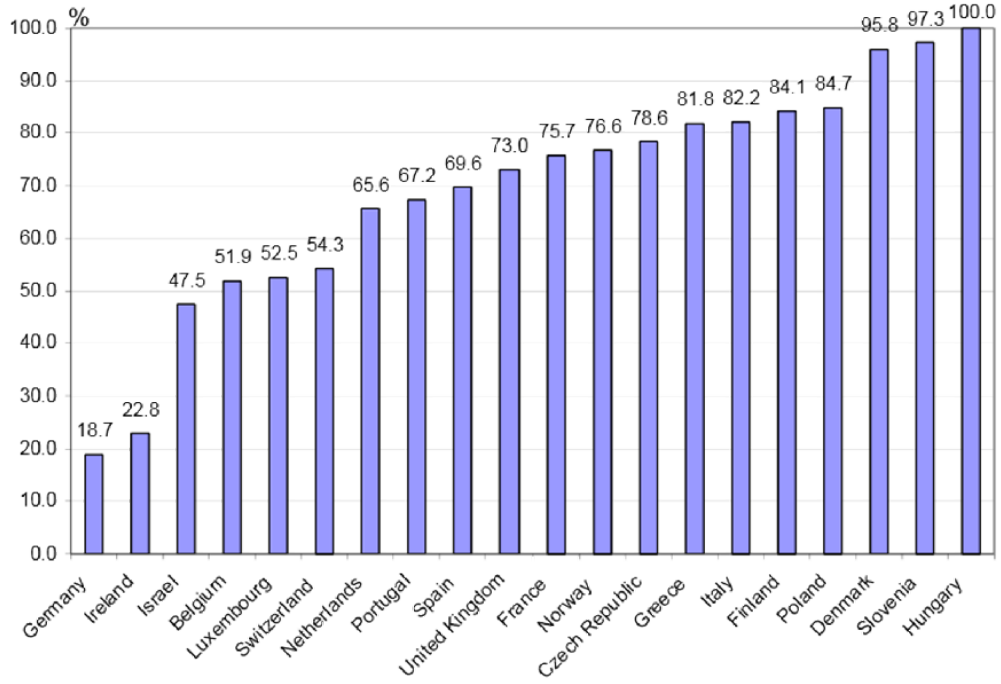

FIGURE 20: PERCENTAGE OF INTERVIEWERS WITH AT LEAST ONE COMPLETED INTERVIEW WITHIN FOUR WEEKS AFTER START OF FIELDWORK, ESS ROUND 1 (SOURCE: KOCH AND BLOHM 2006: FIELDWORK DETAILS IN THE EUROPEAN SOCIAL SURVEY (P. 36))

Information on time stamps can also be used to check on interviewer performance. The time stamp for the first interview conducted by the interviewer after the start of fieldwork provides information about the prioritisation of the survey by the interviewer. The time when the first interview is conducted can also be used as an indicator for interviewer management, e.g. the availability of interviewer staff or if interviewers are working on the study. Koch and Blohm (2006: 36) showed that in some countries (Hungary, Slovenia, Denmark) most of the interviewers conduct an interview within four weeks after the start of fieldwork. In other countries like Germany and Ireland, less than a quarter of the interviewers conducted an interview in the first four weeks of fieldwork (Figure 20).

Combining the information about the time and day of the interview with respondent characteristics can provide an indicator of the inclusion of respondents with certain characteristics in the realized sample. In general we expect that the working population is more difficult to contact during daytime. Analysis by Koch and Blohm (2006: 46) shows that as expected, people in paid work (defined as working more than 1 hour in the last week) are interviewed in every country more often during weekday evenings or on the weekend (Figure 21). This analysis supports the general assumption that the working population is difficult to contact during daytime and is more likely to be contacted either during the evening or on the weekend. It also points to the fact that contacting respondents at different times of the day or on the weekend is important to achieve a representable sample.

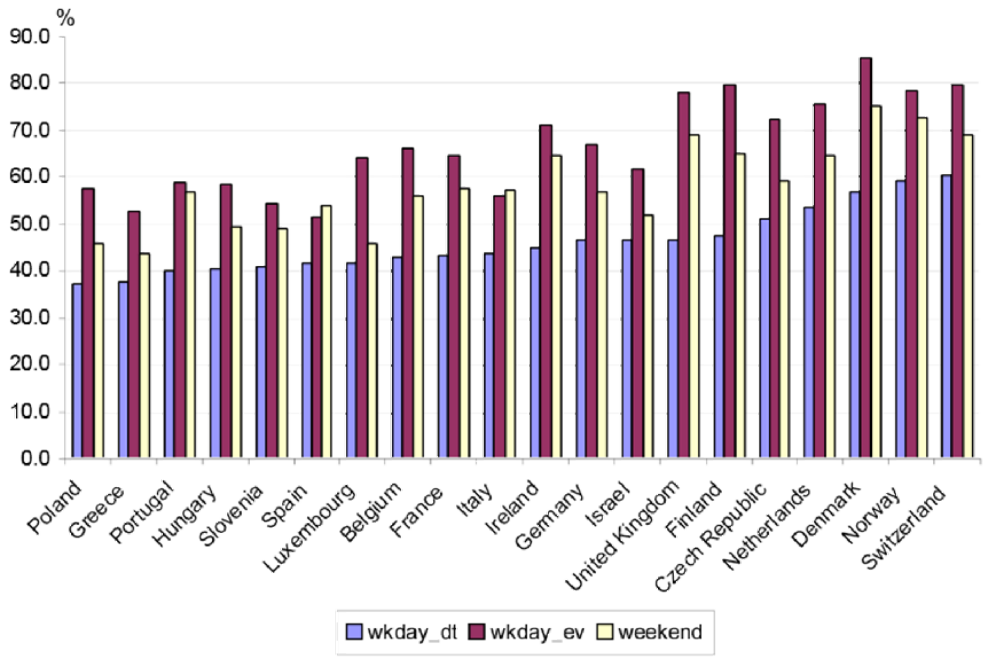

FIGURE 21: PERCENTAGE OF PEOPLE IN PAID WORK IN THE LAST 7 DAYS, BY TIME AND DAY OF INTERVIEW, ESS ROUND 1 (SOURCE: KOCH AND BLOHM 2006: FIELDWORK DETAILS IN THE EUROPEAN SOCIAL SURVEY (P. 46)

Further potential analysis:

- Time and day of interview. Example: At what time are the interviews conducted? Are all interviews conducted in the evening or at the weekend? This can provide information on the time the interviewers are working. If no interviews are conducted at the weekend or in the evening, this might be problematic to reach certain groups of respondents, like the full-time working population. See also Figure 21.
- Average number of days until interviewers completed their first interview (see Koch and Blohm (2006: 34)). This information provides more details than the analysis of the percentage of interviewers with at least one completed interview within four weeks of fieldwork (see figure 21).
- The analysis of time and subgroups can be related to the under- or overrepresentation of groups. Example: Are people in paid work underrepresented on the interviews conducted during the weekdays? See analysis of Koch and Blohm (2006: 48). This can be further analysed by looking at refusals and refusal conversions. Analysis may identify best timing patterns for respondents with specific characteristics.
- Analysing contact sequences from call record data (e.g. obtained from the ESS contact forms) might provide additional information on underrepresented subgroups and indicators for nonresponse. See Kreuter and Kohler (2009).
- Keystroke data provide manifold ways to look at data quality, especially when augmented with further information. Therefore the development of theoretically meaningful indicators is needed. This procedure is described for example in Jans et al. (2013), chapter 9 of Kreuter (2013). Regarding the use of keystrokes for creating key performance indicators (KPIs), they highlight the four indicators interview duration, interview pace, item nonresponse rate and average rate of items missing.

4.4 Cross-cultural and cross-survey analysis

From comparing SHARE's analyses with results from the ESS, we see that findings show some similarities across surveys. We compare the mean interview length between SHARE wave 5 (2013) and ESS round 6 (2012) for the countries that participated in both surveys (see Figure 22). Durations are standardized at the survey's overall mean to better compare interview durations across surveys. We see that in both surveys the same countries tend to have longer or shorter interview lengths. Slovenia, Israel and Spain are the countries with the shortest interview length in SHARE as well as in the ESS. At the other end of the spectrum Germany, Sweden, Denmark and the Czech Republic are among the countries with comparatively long interviews. The majority of the countries show a similar pattern across the two surveys: Spain, Belgium-nl, Switzerland, France, Estonia, Netherlands, Denmark and Germany are close to the dashed line in Figure 22 – which means that they have a similar standardized length in ESS and SHARE.

The comparison of the two surveys suggests that the length of the interview is not a characteristic of the survey only, but also country specific. This might reflect language differences or survey culture differences and goes beyond survey management or the survey topic.

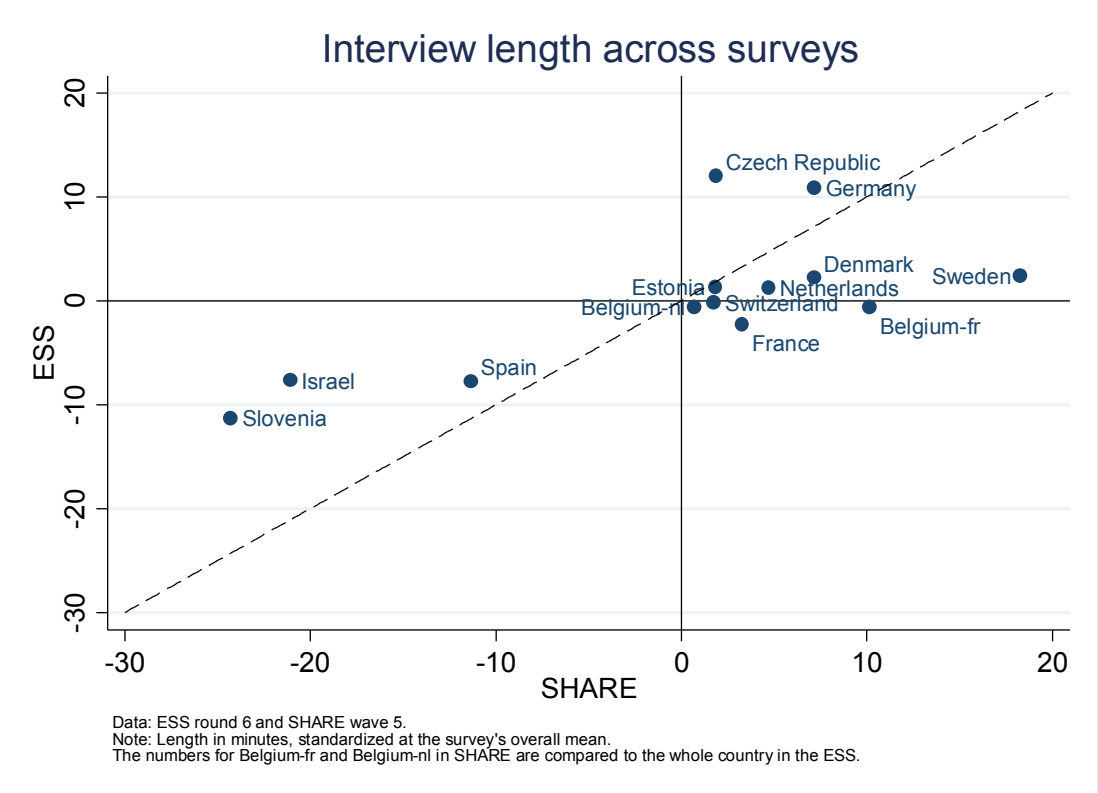

FIGURE 22: INTERVIEW LENGTH FOR ESS ROUND 6 AND SHARE WAVE 5, COUNTRIES PARTICIPATING IN BOTH SURVEYS

5 Conclusion and implications for fieldwork

This report discusses keystroke analyses in SHARE and the ESS and the implications for field work. Keystroke and time stamp data as part of paradata are key data for analysing data quality in survey production. Keystroke data and paradata are useful tools for informing survey management in several areas and throughout various steps of the survey process. Information about the time used provides additional insight into survey data to inform survey managers. We report on analyses conducted during the survey lifecycle of SHARE and after the survey fieldwork in ESS and derive suggestions for further potential analysis. Thereby, this report provides an overview of keystroke and time stamp analysis and its use for fieldwork decisions.

Data collection of paradata and data quality is an important but little discussed topic. Automatic capturing of time provides much more possibilities than manual collection in PAPI questionnaire. Information collected in PAPI can be prone to rounding errors since most interviewers give estimates on time, which can be seen by peaks at 30, 45 or 60 minutes. Automatic capturing of time eliminates the rounding error. But still it is far from error free. The preparation of the raw data and also the analysis requires advanced skills of the data analysts and researchers. Data preparation from the raw keystroke data to the final data involves many decisions for editing and cleaning. Keystroke and time stamp data are available for SHARE and ESS data. The SHARE data are comprehensive and available for project-specific use upon request. The ESS data are less extensive, but publicly available on the ESS website.

The use of keystroke and time stamp analysis can be valuable before, during and after fieldwork. In the *pretest* phase, the data can be analysed in manifold ways to inform decision-making on questionnaire changes. The length of the questionnaire imposes a burden on the respondent. Also longer surveys usually imply higher survey costs. Analysis on the length for the panel and the refreshment samples in SHARE monitor the development of interview length over waves and how changes to the questionnaire might affect the overall interview length. In combination with item non-response, analyses based on pretest data can support decisions on the inclusion, change or exclusion of items. Time measures, especially on item level, are a useful tool for informing questionnaire development.

Analysing paradata *during fieldwork* provides valuable insights on interviewer performance on a regular basis which can be fed back to the survey agencies and interviewers. The focus of ESS and SHARE as cross-national survey is on the cross-country perspective. Parts of the length variation are due to linguistic differences between countries, other differences might be due to non-standardised interviewing. Different survey management strategies and interviewer training styles might cause unwanted variation in international data collection.

Fieldwork monitoring in SHARE comprises a broad range of indicators on different levels. They range from a broad perspective of interview length across all countries to a very detailed level of investigation of item length per interviewer. Length analyses by interviewer can be a good tool to check for interview abnormalities. Interviewers with very short or long average durations do not seem to follow standardised interview behaviour. Further investigation on very short interviews might provide an indication of interview fraud in the total interview, or also of response styles of the respondents like satisficing or straight lining. Looking at the duration of modules and questions can guide survey managers to cases which need further investigation, e.g. data quality checks on the actual survey data. The use of paradata can help to estimate if, for example, introductory texts are read in full length. This information can be used for interviewer training and checks on standardized interviewing. It is important to take the interviewer into account and to disentangle the influence of respondent and interviewer on response styles and overall survey quality. Paradata on interview length and item length can help to analyse the role of the interviewer in the interview process and help to investigate data quality.

Results from the *post survey* checks can be used for informed questionnaire development in future surveys. It can provide insights in the fieldwork process and can be used for cross-national and cross-survey comparison. Besides cultural and linguistic differences different length of the survey or modules can be an indicator for difficulty of a topic or the cognitive ability of the respondent. Combined with respondent characteristics we could show that education and nationality correlate with interview length. Post-survey analysis of suspiciously short interviews combined with measures like item-nonresponse, speeding and shortcutting is a valuable tool for data cleaning.

Looking at the time and date of the interview can provide insights into fieldwork process. For example the number of interviews conducted on average per interviewer, as well as the percentage of interviewers who conducted the first interview within the first weeks of fieldwork, can be used as an indicator of survey management decisions. Also the number of interviews conducted during weekday daytime, during the evenings or on the weekend allows us to learn more about the fieldwork process. Augmented with respondent characteristics, like the occupational status, we can learn more about hard-to-reach populations that are usually underrepresented in surveys and how they are best contacted. This information of course is not limited to ESS and SHARE, but can also be applied to other surveys.

Lessons learned

Keystroke and time stamp analysis is a new field of analysis which offers lots of opportunities. Information about the time can be used in multiple phases of the survey lifecycle to inform survey managers. Paradata are a valuable tool for data quality analysis before fieldwork for developing the questionnaire, during fieldwork to check the data

quality and also for post-survey quality analysis of the interview process. We suggest analysing time stamps and keystroke as part of the quality control process of a survey.

Paradata such as keystroke and time stamp data are raw data. Hence, time for preparation, data cleaning and outlier diagnostic is needed. The automatic recording of time measures in ESS-CAPI and SHARE avoids rounding error caused by interviewers. In general, it is important to evaluate the quality of the original paradata first, that will then be used to analyse data quality of the survey answers. In general, we recommend using time measures throughout the whole survey lifecycle if possible. Analysing the time recording as part of quality assessment is valuable during pretest, during fieldwork and after the survey. The example of SHARE shows that time stamps offer good indicators for assessing item characteristics at the pretest as well as for monitoring interviewers during fieldwork. ESS analyses showed the added value of using time measures in combination with respondent characteristics. However, due to the vast amount of raw data, the relevant indicators need to be carefully selected. The analysis of time stamps and keystroke data adds interesting new aspects on the quality assessment of surveys.

The analysis of time measures for cross-national flagship surveys like the ESS and SHARE revealed some similarities that seem to go beyond survey-specific peculiarities. In our comparisons on interview length we identified a similar cross-national pattern for both surveys. Therefore, linguistic and country-specific influences need to be taken into account when using time measures for data quality assessment or fieldwork monitoring. Information on fieldwork can be used to guide survey researchers in the planning of surveys. Insights from fieldwork analysis can be used to provide guidelines for other surveys which do not yet provide or use keystroke and time stamp data themselves.

Acronyms and Abbreviations

CAPI – Computer-assisted personal interviewing

CATI – Computer-assisted telephone interviewing

CITY – City University London

DASISH – Data Service Infrastructure for the Social Sciences and Humanities

ESS – European Social Survey

GESIS – GESIS- Leibniz-Institut für Sozialwissenschaften

MPG-MEA – Max Planck Society – Munich Center for the Economics of Aging

PAPI – Paper and Pencil Interviewing

SHARE – The Survey of Health, Ageing and Retirement in Europe

SSH – Social Sciences and Humanities

WP – Work Package

References

Börsch-Supan, A. and H. Jürges (2005). The Survey of Health, Ageing and Retirement in Europe – Methodology. Mannheim, MEA.

Bristle, J., M. Celidoni, C. Dal Bianco and G. Weber (2014). "The contribution of paradata to panel cooperation in SHARE." SHARE Working Paper Series **19**.

Couper, M. P. (1998). "Measuring survey quality in a CASIC environment " Proceedings of the Section on Survey Research Methods Section, American Statistical Association: 41-49.

Couper, M. P. and L. Lyberg (2005). "The Use of Paradata in Survey Research." Proceedings of the 55th Session of the International Statistical Institute.

Couper, M. P. and E. Singer (2013). "Informed Consent for Web Paradata Use." Survey Research Methods **7**(1): 57-67.

European Social Survey (2011). Round 6 Specifications for Participating Countries. London, Centre for Comparative Social Surveys, City University London.

Galesic, M. and M. Bosnjak (2009). "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." Public Opinion Quarterly **73**(2): 349-360.

Groves, R. M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau (2004). Survey Methodology. Hoboken, New Jersey, John Wiley & Sons.

Groves, R. M. and S. G. Heeringa (2006). "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." Journal of the Royal Statistical Society **169**(3): 439-457.

Jans, M., R. Sirkis and D. Morgan (2013). Managing data quality indicators with paradata based stastistical quality control tools: the keys to survey performance. Improving Surveys with Paradata. F. Kreuter. Hoboken, New Jersey, John Wiley & Sons.

Jürges, H. (2005). Interview, Module, and Question Length in SHARE. The Survey of Health, Ageing and Retirement in Europe – Methodology. A. Börsch-Supan and H. Jürges. Mannheim, MEA**:** 82-87.

Koch, A. and M. Blohm (2006). "Fieldwork Details in the European Social Survey 2002/2003." ZUMA Nachrichten Spezial **12**: 21-52.

Korbmacher, J. M. and M. Schroeder (2013). "Consent when Linking Survey Data with Administrative Records: The Role of the Interviewer." Survey Research Methods **7**(2): 115-131.

Kreuter, F. (2013). Improving Surveys with Paradata: Analytic Uses of Process Information. Hoboken, New Jersey, John Wiley & Sons.

Kreuter, F. and C. Casas-Cordero (2010). "Paradata." RatSWD Working Paper Series **No. 136**.

Kreuter, F. and U. Kohler (2009). "Analysing contact sequences in call record data. Potential and limitation of sequence indicators for nonresponse adjustment in the European Social Survey." Journal of Official Statistics **25**(2): 203-226.

Krosnick, J. A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." Applied Cognitive Psychology **5**: 213-236.

Loosveldt, G. and K. Beullens (2013). "'How long will it take?' An analysis of interview length in the fifth round of the European Social Survey." Survey Research Methods **7**(2): 69-78.

Loosveldt, G. and K. Beullens (2013). "The impact of respondents and interviewers on interview speed in face-to-face interviews." Social Science Research **42**(6): 1422–1430.

Lynn, P. (2013). "Longer Interviews May Not Affect Subsequent Survey Participation Propensity." Understanding Society Working Paper Series **2013 – 07**.

Malter, F. (2013). "Fieldwork Monitoring in the Survey of Health, Ageing and Retirement in Europe (SHARE)." Survey Methods: Insights from the Field: Retrieved from http://surveyinsights.org/?p=1974.

Malter, F. and A. Börsch-Supan (2013). SHARE Wave 4: Innovations & Methodology. Munich, MEA, Max Planck Institute for Social Law and Social Policy.

Olson, K. and I. Bilgen (2011). "The role of interviewer experience on acquiescence." Public Opinion Quarterly **75**(1): 99-114.

Schmidutz, D. and J. Bristle (2014). Exemplary Analyses of Confidential Paradata: Ethical and Legal Considerations DASISH, Work Package 6, Deliverable 6.3. Retrieved from http://dasish.eu/deliverables/.

Schröder, M. (2011). Retrospective data collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE methodology. Mannheim, MEA.

Tourangeau, R., L. J. Rips and K. Rasinski (2000). The psychology of survey response Cambridge.

Yan, T. and K. Olson (2013). Analyzing Paradata to Investigate Measurement Error. Improving Surveys with Paradata. Analytic Uses of Process Information. F. Kreuter. Hoboken, New Jersey, John Wiley & Sons.