SHARE
Survey of Health, Ageing
and Retirement in Europe
50+ in Europe

Do Interviewers' Reading Behaviors Influence Survey Outcomes? Evidence from a Cross-National Setting

Michael Bergmann, Johanna Bristle

Working Paper Series 23-2016

SHARE • Survey of Health, Ageing and Retirement in Europe • www.share-project.org

European Research Area
SIXTH FRAMEWORK PROGRAMME
SEVENTH FRAMEWORK PROGRAMME
Federal Ministry
of Education
and Research

Do Interviewers' Reading Behaviors Influence Survey Outcomes? Evidence from a Cross-National Setting

Michael Bergmann and Johanna Bristle[*]

Munich Center for the Economics of Aging (MEA),
Max Planck Institute for Social Law and Social Policy

Abstract: Interviewers play a fundamental role in collecting high-quality data in face-to-face surveys. Here, standardized interviewing is the gold standard in quantitative data collection, while deviations from this interviewing technique are supposed to have negative implications for survey outcomes. This paper contributes to the literature on deviant interviewer behavior by analyzing to what extent interviewers vary in their reading behavior with regard to intra-individual changes across the survey's field period and if this has implications for the survey outcomes. Using item-level paradata from the Survey of Health, Ageing and Retirement in Europe (SHARE), we focus our analyses on introduction items to selected modules of the questionnaire. In contrast to previous research, this offers us the possibility to disentangle reading and response time from interviewers and respondents. In addition, the data source allows us to carefully control for confounding effects. Based on a fixed effects approach, our results show systematic changes in interviewers' reading times. First, interviewers' reading times significantly decrease over the survey's field period, even when controlling for period effects, relevant respondent characteristics as well as specific aspects of the interview. Second, a cross-national comparison including 14 European countries plus Israel reveals that the decrease is uniform in nearly all of them suggesting its generalizability over a wide spectrum of conditions. Third, the decrease influences survey outcomes less negatively than expected and to a varying degree depending on the informational content of the item read by the interviewer. However, it is especially relevant for within-survey requests. On the basis of these findings, we discuss possible consequences for questionnaire design as well as interviewer training and fieldwork monitoring.

Keywords: interviewer behavior, paradata, reading time, cross-national survey

*Corresponding author: Johanna Bristle; e-mail: bristle@mea.mpisoc.mpg.de.

Acknowledgment: This paper uses data from SHARE Wave 5 release 1.0.0, as of March 31st 2015 (DOI: 10.6103/SHARE.w5.100) or SHARE Wave 4 release 1.1.1, as of March 28th 2013 (DOI: 10.6103/SHARE.w4.111) or SHARE Waves 1 and 2 release 2.6.0, as of November 29th 2013 (DOI: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260) or SHARELIFE release 1.0.0, as of November 24th 2010 (DOI: 10.6103/SHARE.w3.100). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions).

1. Introduction

The current gold standard in quantitative data collection to minimize interviewer influence is standardized interviewing. The intention of standardized interviewing is to keep the interview situation for all respondents constant in order to produce consistent measurements and to ensure replicability (Groves, et al., 2009). Such characteristics of collected responses are especially important when it comes to large cross-national surveys that are very often conducted face-to-face. This, however, is exactly when the reading behavior of interviewers, as one key component of standardized interviewing, comes into play. Despite the amount of research documenting interviewer effects in general on the measurement of survey variables (Hansen, et al., 1961, Kish, 1962, Freeman and Butler, 1976, Collins and Butcher, 1983, Mangione, et al., 1992, Davis, et al., 2010, West, et al., 2013, Schnell and Kreuter, 2005) as well as with respect to realizing contact (Lipps and Benson, 2005, Blom, et al., 2011) or achieving cooperation (Bristle, et al., 2014, Blom, et al., 2011), the literature on utilizing interviewers' mere reading times of questions and instructions to investigate survey outcomes is scarce.

While some papers try to disentangle interviewer, respondent and item influences on response times (Olson and Smyth, 2015, Couper and Kreuter, 2013), this has not been done yet with respect to interviewers' reading times. Although there is evidence that the overall interview length decreases over the survey's field period (Olson and Peytchev, 2007) and that interviewers become somewhat less careful with survey instructions as their experience with the procedures increases (Olson and Bilgen, 2011, Chromy, et al., 2005), the question how this might influence survey data in a substantial way is less clear. In an early work based on telephone interviews Groves and Magilavy (1986) found significant variation of interviewer reading behavior, but rather low correlation with survey outcomes. Other studies focusing on the role of interviewer experience tap their reading behavior only marginally (Olson and Peytchev, 2007, Lipps and

Pollien, 2011, Olson and Bilgen, 2011). Furthermore, previous work in this field is frequently limited to a single study from one country only. Therefore, the central aim of this paper is to shed light on the question, if and how a change in interviewers' reading times and thus a deviation from standardized interviewing matter with respect to substantive findings in a cross-national face-to-face survey. Especially in a cross-national setting involving different cultural contexts, comparability and standardization of interviewing are of utmost importance.

In face-to-face surveys interviewers play a crucial role in data collection as they are directly interacting with the respondents (Hox, 1994). They can either be seen as one type of error source introducing variance to the interview situation or they can be seen as opportunity to assist respondents during the process of answering survey questions. Strictly standardized interviewing intends to reduce the influence of the interviewer as interviewer deviations are seen as an error source, which might bias the data and negatively affect survey data quality (Fowler and Mangione, 1990). Therefore, one explicit instruction for interviewers is to read question texts exactly as they are worded. According to Groves and colleagues (2009, p. 305) "[u]niform wording of questions asked is perhaps the most fundamental and universally supported principle of standardized interviewing“, as it intends to reduce interviewer bias.

In practice, however, things look different. Studies frequently show that interviewers do not follow the rules of standardized interviewing, but instead change their interviewing behavior as they gain more experience – either from prior surveys (Bradburn and Sudman, 1979, Chromy, et al., 2005) or from earlier interviews over the survey's field period (Cannell, et al., 1977, Fowler, 1991, Olson and Peytchev, 2007). From a theoretical point of view such an interviewer behavior can be explained by principal-agent theory (Jensen and Meckling, 1976, see Kosyakova, et al., 2015 for an application in the field of survey methodology). This perspective considers interviewers as rational agents with own intentions that probably differ from the researcher's objectives. A principal-agent (or more specifically, a moral-hazard) problem arises if the principal (here the researcher) cannot monitor every single action taken by the agent (the interviewer) and if the latter has an incentive to behave in a way that the former does not want. Such unintended actions due to asymmetric information can include, for example, skipping specific questions (Kosyakova, et al., 2015), but also simply shortening questions or speeding through sections of the questionnaire (Fowler and Mangione, 1990, Matschinger, et al., 2005, Brüderl, et al., 2012). The latter becomes especially relevant when interviewers are paid per completed interview (instead of being paid per hour), because then interviewers have a strong incentive to conduct an interview in a short time.

Based on these considerations we can formulate three hypotheses that we will test in our study. First, we expect that only a small proportion of interviewers will strictly follow the rules of standardized interviewing, i.e. read question texts exactly as they are worded. Instead, it seems more likely that interviewers will adjust their reading behavior over the survey's field period resulting in a decrease of reading time for particular questions (shortening effect), even when thoroughly controlling for respondents' characteristics and specific aspects of the interview situation. Second, following the argument that it is a rational behavior to shorten interviews given a payment structure by interview instead of by hour, we expect that such a decrease of reading time is uniform across countries and thus can be generalized. Third, we further hypothesize that, depending on the amount of informational content, which is shortened or read out very fast, this will affect measurement and thus substantive findings to a varying extent. A strong decline in reading time as well as a very short reading time should then be associated with less reliable and less informed responses. In particular, this should be the case with respect to within-survey requests. Within-survey requests ask the respondent to consent or agree to the collection of additional data and go beyond the traditional question-answer situation. This includes, for instance, the linkage to administrative data or the conduction of physical tests and bio-measures (Sakshaug, 2013). Here, we expect that respondents need more information to fulfill the required task. Therefore, interviewers are supposed to give additional information to the respondents.

Against this background, the remainder of the paper is structured as follows: In the next chapter, we introduce the data as well as the variables used and describe the empirical strategy for our analyses. Afterwards, the results are reported in two steps: First, we investigate changes in reading time over the survey's field period taking into account country differences by using a large-scale face-to-face survey including 14 European countries plus Israel. Second, we explore if these changes affect collected responses. We conclude with a discussion of implications of our findings for interviewer training and fieldwork management as well as questionnaire design and monitoring.

2. Data and methods

2.1 Data

The present study uses data from wave 5 Release 1.0.0 of the Survey of Health, Ageing and Retirement in Europe (SHARE; Börsch-Supan, et al., 2013, Börsch-Supan, 2015). SHARE is a multidisciplinary and cross-national panel study, which is conducted biannually. SHARE strongly contributes to the understanding of the ageing process by collecting data on health, socioeconomic status and social and family networks from individuals aged 50 and older and their partners. The fifth wave was collected in the year 2013 in 14 European countries (Austria, Belgium, the Czech Republic, Denmark, Estonia, France, Germany, Italy, the Netherlands, Luxembourg, Slovenia, Spain, Sweden, and Switzerland) and Israel. SHARE is an ex-ante harmonized survey with a high degree of central coordination with regard to questionnaire development, interviewer training, fieldwork monitoring, and data cleaning. Country teams, who are in close contact with the central coordination team throughout the whole survey cycle, implement the survey in the respective countries. The mode of the survey for all participating countries is computer-assisted and face-to-face using a centrally developed CAPI system. The household response rate for the refreshment samples varied from 31% to 60% across countries, according to standards set by the American Association for Public Opinion Research (for details see Kneip, et al., 2015). Magnitude and country differences are comparable to other cross-national surveys in Europe (e.g. ESS7, 2015). In wave 5, retention rates for individuals who participated in the previous wave varied between 71% and 89% (for details see Kneip, et al., 2015).

In addition to the released CAPI interview data, we use item level paradata derived from Blaise audit trails, also called keystroke data (Kreuter, 2013). These data are automatically generated and provide a rich source of information on the interview process.[1] From this data source, latent timers were extracted, which measure the exact duration from when the question appears on the screen until the answer is entered by the interviewer. There is evidence that these so called passive time measures are highly correlated with active time measures, where the interviewer manually records reading or response times, and thus can be seen as valid instruments to measure interviewers' reading behavior (Mulligan, et al., 2003). Keystroke data are available for all CAPI interviews. However, we needed to apply some keystroke-specific exclusion criteria. Firstly, item durations below one and above 1000 seconds were excluded. In a second step, outliers were investigated country by country and excluded from the analyses when they were greater than two standard deviations below or above the mean as usually applied in response time analyses (see Couper and Kreuter, 2013, Heerwegh, 2003, Bassili and Fletcher, 1991).

[1] A detailed documentation of keystroke data in SHARE wave 5 can be found in Bristle (2015).

Overall, a total of 65281 interviews were released for SHARE wave 5 Release 1.0.0. For conceptual reasons, we restricted our sample of analysis to complete interviews and excluded interviews conducted in nursing homes as well as complete proxy interviews. After applying the conceptual restrictions, keystroke preparations, and the outlier exclusions for item durations, we obtained a final analysis sample of 62536 interviews that have been conducted by 1582 interviewers in 15 countries, which means that we excluded about four percent of the observations. The number of observations in the results section might vary slightly by analysis due to missing information on the different outcome and control variables.

2.2 Measurement of main variables of interest

In this paper, our main variable of interest is the reading duration of specific items derived from keystroke data. This type of paradata is rarely used for investigating interviewer behavior in large-scale face-to-face surveys since reading and response times of interviewers and respondents are often intertwined. To obtain a clean measure of interviewer behavior we therefore extracted only those items that contained mere introduction or explanation texts to be read by the interviewer and which were not backed up, edited, or re-entered by the interviewer. These items do theoretically not involve verbal interaction with the respondent. Overall, there are 23 of these introduction items in the SHARE wave 5 questionnaire of which we use six items for analyzing changes over the survey's field period and consequences for survey outcomes. These six items were selected because they contain relevant information for the respondents to provide a well-informed answer. We differ between three types of informational content that appear relevant for a correct understanding of the question or survey request: (1) providing the respondent with information on confidentiality concerns, (2) reading aloud precise definitions of rather difficult or unfamiliar terms, and (3) giving detailed instructions for conducting a specific task. The exact wording of these selected items can be found in Appendix 1.

Regarding information on respondents' *confidentiality concerns*, we use two statements from the questionnaire: one about the scientific use of the data including the explicit acceptance of item nonresponse to sensitive questions and another about the respondents' consent to link the data to administrative records. While the former was read to respondents in all countries at the beginning of the interview, the record linkage project is implemented only in Germany and our analyses are therefore restricted to the German subsample. In terms of introductions containing precise *definitions* of difficult or unfamiliar terms, we use an item that introduces questions about respondents' "out of pocket payments" in the last year and, at the same time, gives examples what should be included here. In addition, we analyze changes in the duration of a battery of questions referring to the respondent's feelings about her "local area", which is defined as the geographical region within a 20 minute walk or one kilometer around the respondent's home. Finally, concerning items with detailed *instructions* for the respondent in order to understand how to properly conduct a specific task, we use a cognitive recall test that asks the respondent to name up to ten words in a certain time frame from a previous given list. Furthermore, we explore changes in reading time with respect to an instruction for the physical measurement of strength and endurance in the respondents' legs.

In our data, we observe interviews over an interviewer's workload. In total we look at 1582 interviewers. As can be seen in Table 1, interviewers' workload is on average about 40 completed interviews. Country-variation is considerable with a mean number of interviews ranging from 25 interviews per interviewer in France to 81 in Israel. The suggested total number of interviews in SHARE (which is 50 interviews per interviewer) is exceeded by more than 25 percent of all interviewers. However, 95 percent of the interviewers are within 104 interviews, which is why we restrict the following analyses to interview sequences up to 100 interviews. The selected introduction items that contain confidential information show a mean reading time of almost 13 seconds for the overall intro to the SHARE interview and about 80 seconds concerning the request for record linkage. The two selected items including definitions have a mean duration of about 14 (intro to health care expenditures) and 12 seconds (intro to local area information), respectively. Items containing instructions take on average about 26 (intro to recall test) and 24 seconds (intro to chair stand measurement).

Table 1: Distribution of main variables of interest

	N	Mean	Std. Dev.	Percentiles				
				p5	**p25**	**p50**	**p75**	**p95**
Total number of interviews	62563	39.55	32.68	4	16	31	54	104
Intro to overall SHARE interview	36581	12.71	16.99	1	2	7	17	44
Intro to record linkage	4822	79.56	75.82	2	23	57	112	242
Intro to health care expenditures	57647	13.77	15.01	1	2	8	23	41
Intro to local area information	39005	12.01	10.47	1	3	9	19	32
Intro to recall test	59839	26.31	24.22	1	5	21	38	75
Intro to chair stand	57123	24.16	31.72	1	2	7	37	94

Cell entries are total number of interviews in the first row and seconds otherwise (all unweighted).

Our empirical analyses are organized in two steps: In the first step, we investigate changes in interviewers' reading duration over time. As said before, we hypothesize a shortening effect that leads to a decrease in reading times towards the end of the field period that is uniform across countries. In the second step, we analyze the impact of this shortening behavior on survey outcomes that should be affected by reading these introduction texts thoroughly or not. Table 2 gives a short overview on the used introductions and the corresponding survey outcomes. The exact wording and answer categories of the survey outcomes can be found in Appendix 2.

Table 2: List of used introduction variables and the respective survey outcomes

Type of content	Reading item	Survey outcome
Confidentiality	Intro to overall SHARE interview	Nonresponse to income question
Confidentiality	Intro to record linkage	Consent given to record linage
Definition	Intro to health care expenditures	Payed anything out-of-pocket
Definition	Intro local area information	Feeling part Cleanliness Help available Vandalism or crime
Instruction	Intro to recall test	Amount of words recalled
Instruction	Intro to chair stand	Compliance with chair stand test

Concerning items about confidentiality concerns, we suppose that reading this information not thoroughly will result in a lower amount of item nonresponse to a sensitive question, such as the respondents' income, as they are not aware of its acceptance. Moreover, we expect a lower consent rate to record linkage if the interviewers superficially speed through the statement of the linkage procedure because respondents might feel uncomfortable about their privacy.

Regarding items with a definition of uncommon terms, we expect that omitting or speeding through considerable time consuming explanations reduces the likelihood to report out of pocket payments for doctoral visits. Similarly, we expect that shorting the definition of "local area" (meaning the geographical region within a 20 minute walk or a one kilometer around the respondents' home) affects residential satisfaction, i.e. their personal affiliation, security, upkeep, and social relations in the neighborhood.

Finally, with respect to our last item type with informational content, we expect respondents to perform better who received detailed instructions and who were allowed enough time to comprehend the instructions before starting the test. Good performance is measured here in terms of more words recalled indicating the effort respondents put into performing the test thoroughly. The second test is a physical measurement for strength and endurance in legs called chair stand, where the respondent is asked to stand up from a sitting position in a certain way. Here as well, missing instructions from the interviewer's side is supposed to negatively influence the compliance of the test. A description of the survey outcomes can be found in Appendix 3.1.

2.3 Empirical model

Throughout our analyses, we use fixed effects regression models (Allison, 2009, Wooldridge, 2013) to control for time invariant differences between interviewers and to focus on intra-individual changes of interviewers' reading behaviors over the survey's field period.[2] Formally, we can write our general model based on the terminology of Wooldridge (2013) in the following way:

$$y_{it} = \beta x_{it} + \alpha_i + u_{it} \qquad (2.3.1),$$

where y_{it} is the reading behavior for a set of interviewers (i = 1, …, n) that is measured at several points in time (t = 1, …, T representing each conducted interview). Further, x_{it} denotes a vector set of explanatory variables (see Appendix 3.2 for a full list), while β is the corresponding vector of coefficients. The error term is split into two components: While α_i varies only across individuals and is stable over time, u_{it} represents random variation at each point in time.

In general, we expect the reading behavior to change over the field period. More specifically, and as hypothesized above, we expect a sharp decrease in interviewers' reading time during the first interviews due to increasing experience and a more stable pattern later on. Therefore, we model the number of interviews as our main explanatory variable in a semiparametric way by using piecewise regression lines, called linear splines (for an introduction, see Keele, 2008). This approach allows for local approximations over the number of

[2] In the literature on interviewer effects, multi-level analyses are more common. In contrast to a multi-level approach, the fixed-effects approach, however, has the advantage that time-constant unobserved variables are allowed to be correlated with the observed variables of interest. Because such associations can heavily bias our conclusions about interviewer behavior, we believe that the disadvantage of larger standard errors in fixed effects models will be more than compensated.

interviews, while at the same time it is more flexible than a quadratic modeling and easier to interpret than polynomials. In practice, this means that several dummy variables enter the equation and specify the local areas of the continuous variable of interest. We set spline knots at the 2nd, the 10th, and the 50th interview based on the following considerations: Between the first and the second interview the decline should be steepest as this resembles a real-scenario test. Besides trainings, this might be the first time the interview is conducted completely with a real respondent in that wave. Therefore, the reading time in the first interview might be higher than what we would consider a properly conducted, standardized interview. From a substantial point of view, the most interesting time interval thus is between the 2nd and the 10th interview. In this period we expect interviewers to adjust their reading behavior. Here, the standardized interviewing technique as trained during the interviewer training is successively replaced by increasing experience and the interviewing behavior is expected to become less standardized. After some interviews, we expect that behavioral adjustments have taken place more or less leading to a weakening of the decrease in reading time. The knot at 50 interviews was chosen, because it indicates the advised maximum number of interviews in SHARE.

When analyzing changes in interviewers' reading behaviors over time the error term α_i drops out and possible correlations with x_i do not bias our estimates. The same is true for all time-invariant explanatory variables included in x_i. Thus, the equation we use in the following can be rewritten as:

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i \qquad (2.3.2),$$

where $y_{it} - \bar{y}_i$ is the time-demeaned data on the reading behavior y, and similarly for $x_{it} - \bar{x}_i$ and $u_{it} - \bar{u}_i$. Basically, this fixed effects transformation has the effect that we can neglect all stable interviewer characteristics that shape the general reading behavior, such as their individual base speed of reading, their language, their cognitive abilities, as well as socio-demographic characteristics of the interviewer, whether or not they have been measured. However, in order to draw reasonable conclusions concerning the effect of an increasing number of conducted interviews on interviewers' reading time, we need to carefully control for potential time-variant confounders that might influence the interviewer's reading behavior besides the number of conducted interviews up to this point.

Most prominent with respect to control variables are respondent characteristics that lead to an adjustment on the part of the interviewer, such as the age of the respondent, subjective and objective health indicators, years of education, gender, the current job situation, household size, living in an urban area or not, as well as engagement in social activities, and the respondents' previous interview experiences. Together with specific aspects of a certain interview, such as if there is a partner living in the household or not, if it is a baseline or a panel interview, if the interview is the first or second in the household, if the interview took place during the last two months of the survey's field period, and how many interviews the interviewer already conducted on a specific day, these properties characterize the concrete interview situation that changes from one interview to the next. Not taking this into account bears the risk of attributing shorter reading time to interviewer behavior, although this effect might be confounded with changes of the sample composition. For the same reason, we also control for period effects that cover changes in the variables under consideration by adjusting for the number of days in the field. Finally, we weight our data to avoid that interviewers with a very low (high) number of interviews exert an over-proportional small (large) influence on the results by giving each interviewer an equal importance.

Against this background, we run linear fixed effects regressions explaining changes in the interviewer's reading behavior as dependent variable in the first step. The fixed effects approach enables us to disentangle the mechanisms potentially leading to a decline in reading time. After carefully controlling for potential time-variant confounders and explicitly estimating change in reading behavior using splines, we are confidant to ascribe a negative effect of the increasing number of conducted interviews, especially between the second and the tenth interview, to an interviewer's learning behavior to shorten, skip or speed through the item.

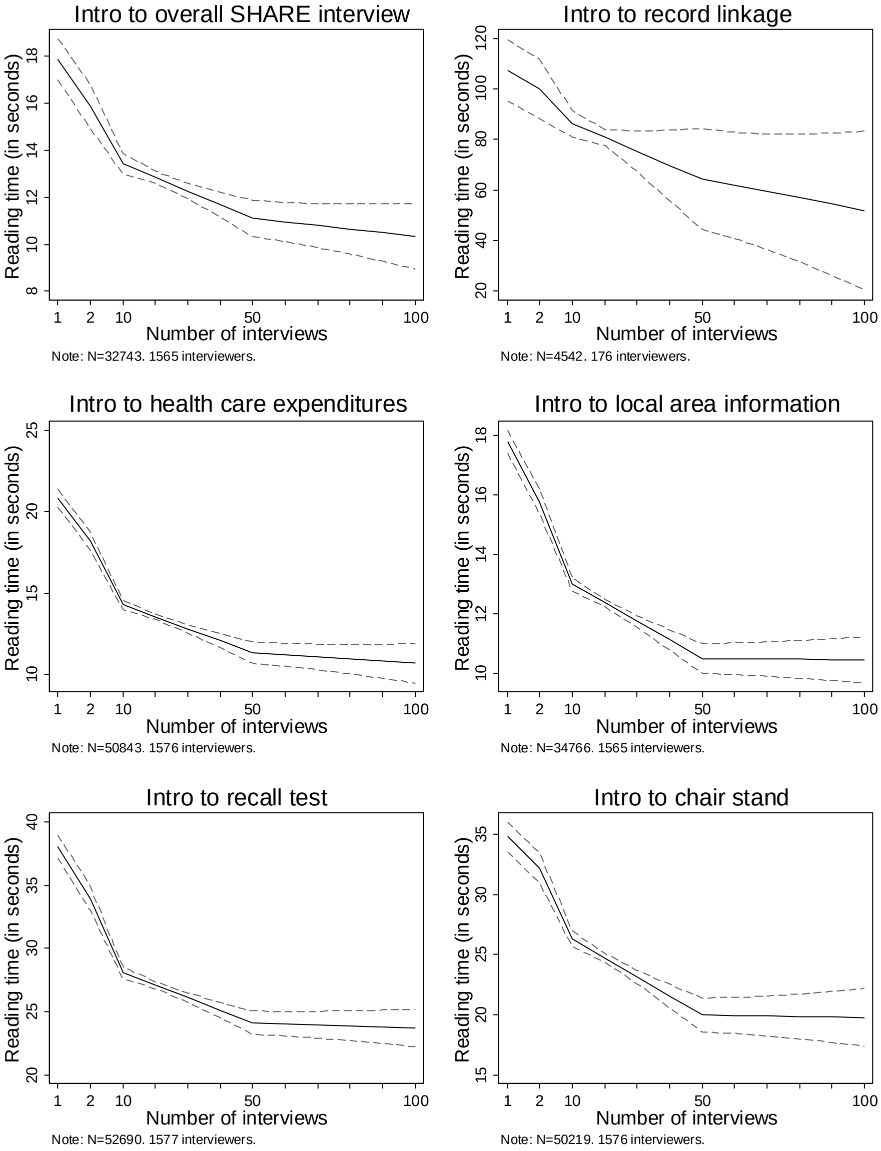
In the second step of our empirical analyses, our main variable of interest, i.e. the changing interviewer's reading duration, acts as independent variable in order to explain substantial consequences on the above mentioned survey outcomes that then serve as dependent variables. Otherwise, our general model described in equation 2.3.2 stays the same. Thus, we again run linear fixed effects regression models and control for period effects, confounding effects of the changing sample composition including respondent characteristics, and aspects of the specific interview situation that might differ between interviews.

3. Results

3.1 Change of interviewers' reading behavior over the survey's field period

In order to analyze interviewer's reading behavior over the survey's field period, the fixed effects approach we use allows us to look at changes in reading time within one interviewer over the course of her workload and to implicitly control for all time-invariant confounders. After controlling for period effects and the interview situation we obtain a reliable trend that is shown in Figure 1. Here, predicted reading time is displayed over the number of completed interviews with 95% confidence intervals. To explicitly model the non-linear function of the development, we use splines as described above. On the x-axis, the labels indicate that spline knots are set at 2, 10, 50, and 100 interviews. Please note that the x-axis is stretched in the beginning to highlight the spline structure. In fact, the decline in this interval is even steeper than it appears in the graph.

Figure 1: Change in reading behavior over number of interviews


Intro to overall SHARE interview

Note: N=32743. 1565 interviewers.

Intro to record linkage

Note: N=4542. 176 interviewers.

Intro to health care expenditures

Note: N=50843. 1576 interviewers.

Intro to local area information

Note: N=34766. 1565 interviewers.

Intro to recall test

Note: N=52690. 1577 interviewers.

Intro to chair stand

Note: N=50219. 1576 interviewers.

Data: SHARE wave 5.Fixed effects regressions with spline knots at 2, 10, and 50 interviews. Controlled for days in field, respondent characteristics, and interview specifics. Weighted results.

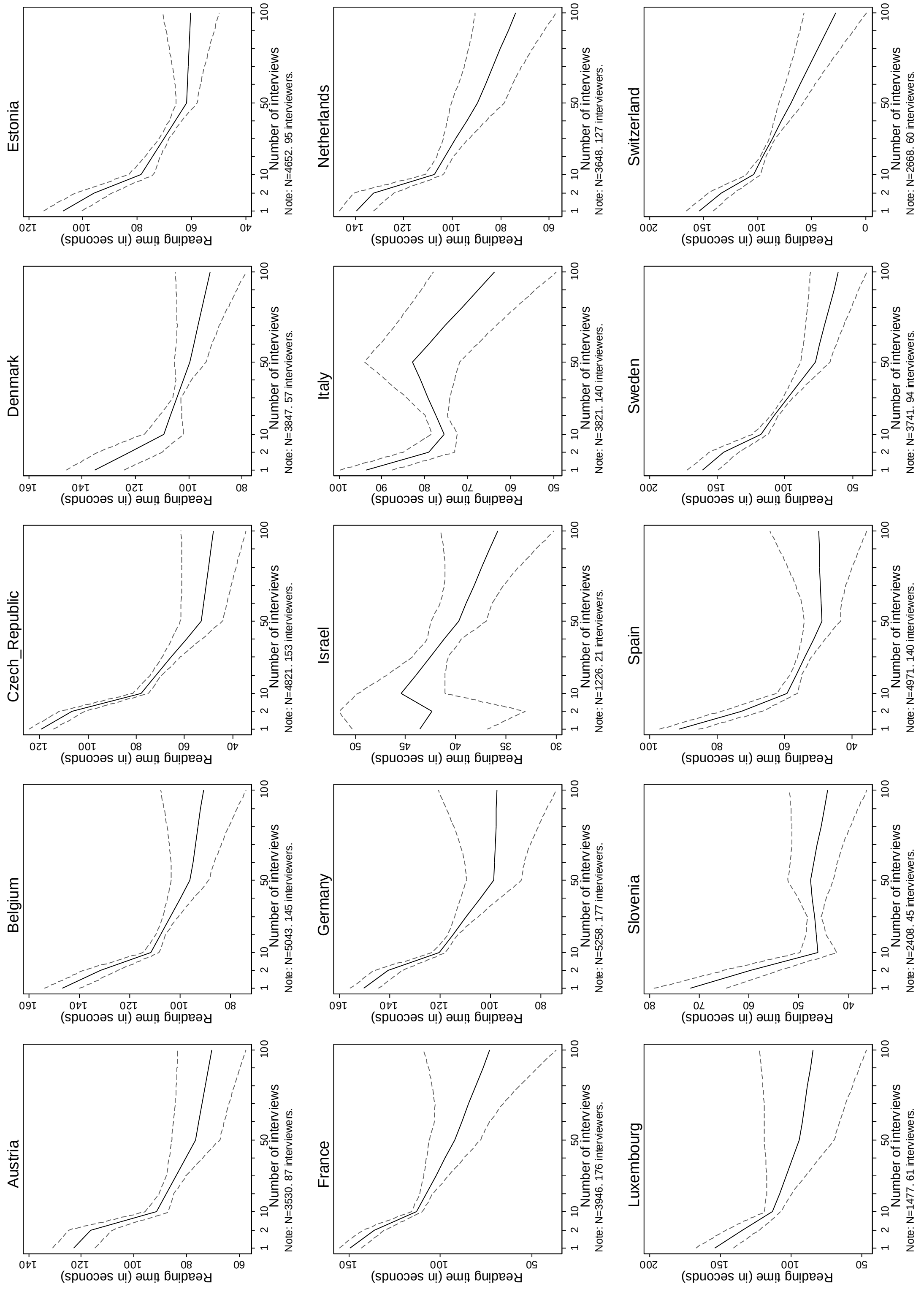
Overall, we see a strong decline for all six introduction items of SHARE. The steepest decrease can be observed in the beginning. For instance, the introduction to the SHARE interview has a mean reading time of about 18 seconds at the first interview and drops to about 13 seconds at the tenth interview, which is a decrease in reading time of almost 30 percent. Until the 50th interview, the decrease is still significant but attenuated a bit. Between the 50th and the 100th interview all six observed items only show marginal decreases, while at the same time standard errors increase due to the lower number of interviewers having such a large workload.[3] Because it is plausible that between the first and the second interview external factors, such as a better preparation of additional materials, play a considerable role, we are most interested in the development in reading time between the 2nd and the 10th interview. A change within this time interval is what we term a learning effect of the interviewer, which is negative and significant for all selected items. This is also the case, when we look at other intro texts, where we do not assume further consequences on survey outcomes (not shown here).

When we allow for country-specific heterogeneity, the overall picture stays more or less the same. Figure 2 shows the cumulated reading time of all intro texts displayed above except the record linkage item[4]. Missing values of certain items have been imputed by the country-specific mean of the reading time at a particular interview in order to obtain a comparable measure. Otherwise, interviewers who have not asked every selected items and thus lack those durations would appear much faster than interviewers who very quickly read out all items to the respondents. Based on this cumulated measure, we again run fixed effects regressions with the same covariates to control for period effects, relevant respondent characteristics, and interview specifics as above, but now distinguish between countries.

[3] All spline coefficients and the respective regression models are reported in Appendix 4.

[4] The record linkage item was not asked in all countries and excluded to avoid redundancies and confusingly small graphs.

Figure 2: Country-specific change in reading behavior over number of interviews



Data: SHARE wave 5.Fixed effects regressions with spline knots at 2, 10, and 50 interviews.
Controlled for days in field, respondent characteristics, and interview specifics. Weighted results.

Overall, we see a comparable strong decrease in interviewers' reading time within the first ten interviews. What attracts attention is the very different (starting) level of interviewers' reading time that can be attributed mostly to language differences. While there are some rather "slow" countries, such as Belgium, France, Germany, Sweden, or Switzerland, there are also countries, such as Israel, Slovenia, and to a lesser degree Spain, in which interviewers are comparatively "fast" right from the very beginning.

Despite this country-specific heterogeneity in interviewers' reading behavior the general pattern of a significant decrease in reading time is rather stable: After a strong decline in the beginning it is attenuated somewhat but still remains negative. There are only three exceptions from this general trend: Israel, Italy, and Slovenia. Here, the decrease is not monotonous. In Israel and Italy, the change between the 2nd and 10th interview is not significant. In Israel and Slovenia this can be explained by the very low level of reading time in the beginning. This simply leaves less room for the interviewers to speed up their reading any further. Additionally, Israel had only very few interviewers in total, which might also influence the findings to a certain degree. In contrast, Italy has a relatively high average reading time and also does not show a significant decline in the middle part of an interviewer's workload. From contact with the country team in Italy we know that the overall interview length as well as reading time for specific items was part of the regular monitoring reports, which have been shared with the interviewers. Therefore, the interviewers know that their behavior is being observed at least partly.

3.2 Robustness checks

We conducted several robustness checks to verify our results (see Appendix 5). In a first specification, we re-ran all our models using a logarithmic transformation of the reading time. This proceeding is sometimes recommended in the literature on response times in order to take into account the typically highly right-skewed distribution of these variables (e.g. Yan and Olson, 2013). However, in our case such a transformation neither improved the model fit nor changed the results in a substantial way (see Appendix 5.1). Therefore, we kept the more intuitive scaling of reading time in seconds.

Second, we want to address the concern that our results are biased due to a non-random selection of interviewers. Therefore, we excluded interviewers with a very low or very high total number of interviews to see if these observations substantially influence our results. This is not the case. After excluding these cases, findings remained virtually the same (see Appendix 5.2).

Along similar lines, our third robustness check addresses the issue that some interviewers drop out quickly and some stay working longer as interviewers for SHARE throughout the field period. In this respect, one can imagine that the amount of interviews conducted by a specific interviewer is not random, but rather due to unobserved characteristics, such as communication skills. This implies a violation of the missing at random assumption (MAR) for unbalanced panels (Rubin, 1976) and limits the generalizability of the fixed effects estimates. Therefore, we created a series of balanced panels and re-estimated our models for these subgroups of interviewers. While learning effects for reading the introduction to the overall SHARE interview and to record linkage are rather unstable across the subgroups, the coefficients for all other introduction texts are very stable and comparable to the results based on the overall sample (see Appendix 5.3).

3.3 Does it matter? Consequences of decreasing reading time for survey outcomes

In the second stage of our analysis, we analyze to what extent the observed decline in reading time substantially affects survey outcomes. Therefore, we use the reading time of the selected introduction texts as main explanatory variables now, while our dependent variables are the survey outcomes described in Table 2 above. We again run fixed effects models and control for period effects, respondent characteristics, and the interview situation the same way as before. The results are presented in Table 3. The coefficients presented in the last column equal the effect of the average change in reading time in the sample. This provides a more sensible interpretation of the effect size than a regression coefficient.[5]

Table 3: Intro-specific regressions on survey outcomes

Reading item	**Survey outcome**	**Avg. change in reading time (in seconds)**	**Effect of avg. change**
Intro to overall SHARE interview	Refusal to income [0;1]	-5.4	-0.000 (0.001)
Intro to record linkage	Consent given [0;1]	-55.1	-0.050*** (0.008)
Intro to health care expenditures	Payed anything out of pocket [0;1]	-12.2	-0.006 (0.003)
Intro to local area information	Feeling part [0;1]	-8.3	0.000 (0.002)
	Cleanliness [0;1]		0.005 (0.003)
	Help available [0;1]		-0.005 (0.003)
	Vandalism or crime [0;1]		0.008* (0.003)
Intro to recall test	Amount of words [0;10]	-18.4	-0.018* (0.008)
Intro to chair stand	Compliance with test [0;1]	-16.9	-0.024*** (0.001)

Note: Each line represents an own linear fixed effects model on the respective survey outcome with reading time of intro text as explanatory variable and days in field, respondents characteristics, and interview specifics as controls.
Panel-robust standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

To give a concrete example: The average decrease in reading out the record linkage intro between the first and the last interview of each interviewer is about 55 seconds in our sample (see column 3). This huge decrease is very likely due to not giving the respondents enough time to carefully read the SHARE statement about privacy concerns. If we want to know what this decrease means for a connected survey outcome, such as the consent rate, we have to multiply this value with the regression coefficient obtained in the fixed effects model. The result of this calculation (-.050) is

[5] See also Appendix 6 for full regression tables including all control variables.

shown in column 4. Thus, the average decrease in reading time with respect to the record linkage item reduces the consent rate by five percentage points.

Similar findings are reported with respect to the decrease in reading out the introduction of the cognitive recall test and the chair stand measurement. Thus, the average decrease in reading out the latter of 16.9 seconds results in a reduced compliance rate of 2.4 percentage points. While these effects explicitly concern within-survey requests, there is also a significant effect with respect to an intro item including a definition. Thus, a shorter reading time of the local area definition results in stating a higher rate of vandalism or crime there. However, this effect is rather small and only slightly significant. The other survey outcomes (refusal to the income question, out of pocket payments, most local area evaluations) are not changed significantly by a decreased duration of the introduction item. In the case of the virtually unchanged refusal rate to the income question, this might be due to the fact that there are two opposing effects cancelling each other out. Thus, on the one hand, some respondents indeed might show the assumed lower refusal rate, because they have not been explicitly told by the interviewer that such a behavior is acceptable. On the other hand, many respondents might react the opposite way and deny a substantial answer to the income question, because they have concerns about their privacy.

When looking at country-specific models, the main results are resembled for nearly all survey outcomes under consideration. Only few countries show minor deviations and these are far from being large or systematic. Therefore, we do not show these models to avoid redundancies. Overall, our analyses again are rather stable in a cross-national perspective and generally support our interpretation of findings described above. Furthermore, in some instances observed differences between countries are quite plausible as the underlying conditions differ substantially. This can be illustrated, for example, in the case of out of pocket payments. Here, the underlying health care system and the national average of out of pocket payment play a crucial role. If almost everyone or almost no one pays out of pocket for health care, reading out the definition does not matter. However, in countries with an average level of out of pocket payment between 25 and 75 percent, providing the respondent with a proper definition indeed matters. Here, the observed decrease of the introduction text shows a significant effect (see Appendix 7).

So far, we restricted our analyses to effects of (within) changes in reading time. Although it is crucial to assess if there is a causal effect of decreasing reading time on survey outcomes, the absolute level of the survey outcomes is not easy to interpret. Therefore, in the following we complement this within perspective by looking at absolute differences between interviewers and how these differences affect the overall level of survey outcomes. This gives a more intuitive feeling for the observed deviations from standardized interviewing by speeding through the questionnaire or even skipping introduction texts. In Table 4 we report the predicted outcomes for two different levels of reading time. The first scenario in column 2 shows the stated value of the respective survey outcome if all interviewers would read the introduction text at the 10%-level, i.e. read the question relatively fast. Then we look at a scenario with interviewers reading out the introduction texts in a standardized way, which is assumed to take rather long. Here, we set the value at the 90%-level of the distribution.

Table 4: Predicted outcomes for different levels of reading time

	If all interviewers would read the intro at the … of the distribution	
	10%-level ("fast")	**90%-level ("slow")**
Refusal to income [0;1]	0.085	0.085
Consent given [0;1]	0.530	0.653^{***}
Payed anything out of pocket [0;1]	0.455	0.471
Feeling part [0; 1]	0.921	0.920
Cleanliness [0; 1]	0.861	0.846
Help available [0; 1]	0.862	0.876
Vandalism or crime [0; 1]	0.197	0.173^{*}
Amount of words [0; 10]	5.337	5.393^{*}
Compliance with test [0;1]	0.819	0.923^{***}

Note: Predicted values are based on marginal effects after running item-specific linear fixed effects models on survey outcomes.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Substantial differences in outcomes can be observed again for consent to record linkage and compliance to the chair stand test. To give a concrete example: If all interviewers would take one second (10%-level) on the introduction item to the chair stand test, 81.9 percent of the respondents would comply. If all interviewers would take 75 seconds (90%-level) instead, the compliance rate would be at 92.3 percent, a difference of 10.4 percentage points. For the other outcomes, the effect is not substantial, although the difference for vandalism and the amount of words recalled are statistically significant at the 5%-level. Overall, our findings suggest that reading time matters especially for within-survey requests and not so much for the other survey outcomes under investigation.

4. Discussion

In face-to-face surveys, standardized interviewing is seen as the gold standard for quantitative data collection intending to keep the interview situation as similar as possible for all respondents and thus ensuring comparability and replicability of the data collected. However, in practice things look different: Our analyses show that, in contrast to the goal of standardized interviewing, interviewers' reading time significantly decreases over the survey's field period. As this pattern is nearly the same in all countries regardless of language differences, our results suggest that strongly shortening or even skipping intro texts to questions is anything but an exception. Of course, a decrease in reading time alone must not be problematic for data quality per se. Therefore, the most important question we tried to answer in this article is: does it matter? Or more concrete, does a decrease in reading time have substantial consequences for survey outcomes? As it is often the case, a precise answer to this question is difficult and depends on the underlying circumstances. However, one important aspect in this respect is the amount of informational content in questions that have been shortened by the interviewers. Thus, our findings reveal a rather large effect of decreasing reading time on within-survey requests, such as the consent given by the respondents to link their survey answers to administrative data or doing a physical test to measure the respondents' endurance. In contrast to previous research, where it has been shown that faster interviews from more experienced interviewers are correlated with higher rates of acquiescence (Olson and Bilgen, 2011), our results reveal a negative effect on consent. While we agree with the assumption that increased pace throughout the interview might create an environment where respondents have less time to think, thereby increasing their likelihood of satisficing, the conclusion that this should lead to more acquiescent answers obviously depends on further conditions, such as the content of the question. Our findings suggest that respondents have a rather keen intuition for not taking their fears seriously – either with respect to confidentiality concerns or regarding their physical abilities. Thus, when confronted with questions that have been read out in a superficial way, they frequently disagree to give their consent to such within-survey requests.

Besides these two rather clear "yes/no" questions, other survey outcomes show much smaller effects, although it seems rather unrealistic that the interviewers read out every piece of information being relevant for the respondents to give reliable and informed answers. For survey researchers this is good news. However, our results should not lead researchers to conclude that strong deviations from standardized interviewing can be neglected completely. Instead, we rather see various implications of our results in different areas that we want to expose in the following. Thus, one major implication of this study concerns questionnaire design. It is well known that the visual presentation of questions and answers can influence the way information is read and also whether it is read at all (e.g. Jenkins and Dillman, 1997, Christian and Dillman, 2004, Dillman, et al., 2014). This, of course, is not only valid for respondents, but also for interviewers. Usually, however, questionnaire design only takes into account respondents, while neglecting the important role interviewers play in this context. To give an example: Introductions are usually presented in block texts. However, such a format requests more concentration to read them out completely and increases the probability that something is left out by mistake or even intentionally. In this respect, a clear structuring concerning the content of introduction texts might be a useful starting point.

Another important implication is related to fieldwork management and monitoring. Overall, the interviews of the study utilized in this paper are relatively long (Bristle, 2015). Therefore, shortening interviews seems to be a rational behavior for interviewers to maximize their utility in this setting. Furthermore, in the study under investigation all survey organizations pay their interviewers per interview. Some have additional interviewer bonuses or reimbursement for travel costs and travel time, but no standard payment-per-hour scheme. In all settings, interviewers have strong incentives to aim at conducting relatively short interviews or speed through specific parts of the questionnaire. As payment-per-interview schemes are usually applied in large-scale surveys, such as SHARE, and survey organizations also have a strong incentive to stay with this routine in terms of minimizing their own management costs, it will be rather difficult to change this. One possible starting point, however, could be to add steeper bonus and/or malus systems that reward good interviewers and punish more delinquent ones. In this respect, it is important to emphasize that good interviewer behavior should not only be measured in terms of response rates, but more closely monitor their actual behavior in the interaction with respondents.

Our results show that interviewers' reading time decreases a lot over the field period. Therefore, we argue that monitoring interviewers more closely is necessary in order to keep their work quality high. For instance, we know from the Italian case that interviewers perform very well when receiving feedback on their interview length and reading time. Closely connected with this, we finally see implications of our findings for interviewer training. Thus, it seems necessary that interviewers, independent from their experience, receive clear instructions stating how important a careful reading of every question is. A re-training might be helpful to reinforce good interviewing practices, especially for interviewers with a very large workload. This is an important aspect, as our analyses show that especially the fastest interviewers are often those with large workloads that at the same time reveal the largest deviances from average with respect to our analyzed survey outcomes.

As with any study, our analysis has limitations. First, although we carefully selected only items that measure sole interviewer behavior, we cannot completely rule out that verbal interactions between respondent and interviewer took place during the measurement of the introduction item. This means the durations measured might be longer than the actual reading time. Taking this into account, the shortening effect we observe then is even more striking. Second, on a similar note, we cannot disentangle what exactly happened during reading out the introduction items. While we only observe interviewers' reading time, we simply do not know if a change in the duration is due to shortening, skipping, or changing words in the introduction item. For this purpose, experimental designs are needed. And third, although we tried to control for several areas of influence, such as respondent characteristics, specific aspects of the interview, and timing effects, we cannot rule out a potential bias due to other relevant variables that we have not taken into account or just are unobservable.

Further research might want to investigate country differences more deeply. The reasons for country-specific deviations in reading time were not investigated further in our analyses. We assume that language differences explain a major part of it, but we have not explicitly tested it. Further research might be able to disentangle influences due to language, survey climate, or different survey management strategies (e.g. interviewer training). Finally, in this study, we focused on behavior changes over the course of a field period, thereby neglecting the long-term panel structure of SHARE. Further research might want to look at changes over different waves as changes in interviewer training or monitoring strategies might also influence interviewers' reading behaviors to some degree.

References

Allison, Paul D. 2009. *Fixed Effects Regression Models*. Thousand Oaks, CA: SAGE Publications.

Bassili, John N., and Joseph F. Fletcher. 1991. "Response-Time Measurement in Survey Research. A Method for Cati and a New Look at Nonattitudes." *Public Opinion Quarterly* 55: 331-46.

Blom, Annelies G., Edith D. De Leeuw, and Joop J. Hox. 2011. "Interviewer Effects on Nonresponse in the European Social Survey." *Journal of Official Statistics* 27: 359-77.

Börsch-Supan, Axel. 2015. Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release Version: 1.0.0. SHARE-ERIC. Data Set. doi: 10.6103/SHARE.w5.100.

Börsch-Supan, Axel, Martina Brandt, Christian Hunkler, Thorsten Kneip, Julie Korbmacher, Frederic Malter, Barbara Schaan, Stephanie Stuck, and Sabrina Zuber. 2013. "Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE)." *International Journal of Epidemiology* 42: 992-1001.

Bradburn, Norman M., and Seymour Sudman. 1979. *Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research*. San Francisco, CA: Jossey-Bass.

Bristle, Johanna. 2015. "Measuring Interview Length with Keystroke Data." In *SHARE Wave 5: Innovations & Methodology*, edited by Frederic Malter and Axel Börsch-Supan, 165-76. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Bristle, Johanna, Martina Celidoni, Chiara Dal Bianco, and Guglielmo Weber. 2014. "The Contribution of Paradata to Panel Cooperation in SHARE." *SHARE Working Paper (19-2014)*. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Brüderl, Joseph, Bernadette Huyer-May, and Claudia Schmiedeberg. 2012. "Interviewer Behavior and the Quality of Social Network Data." In *Survey Standardization and Interviewers' Deviations - Impact, Reasons, Detection and Prevention*, edited by Peter Winker, Natalja Menold and Rolf Porst, 1-14. Frankfurt: Peter Lang.

Cannell, Charles F., Kent H. Marquis, and André Laurent. "A Summary of Studies of Interviewing Methodology." Washington, DC: US Government Printing Office, 1977.

Christian, Leah Melani, and Don A. Dillman. 2004. "The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions." *Public Opinion Quarterly* 68: 57-80.

Chromy, James R., Joe Eyerman, Dawn Odom, Madeline E. McNeeley, and Art Hughes. 2005. "Association between Interviewer Experience and Substance Use Prevalence Rates in Nsduh." In *Evaluating and Improving Methods Used in the National Survey on Drug Use and Health*, edited by Joel Kennet and Josph Gfroerer, 59-87. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Collins, Martin, and Bob Butcher. 1983. "Interviewer and Clustering Effects in an Attitude Survey." *Journal of the Market Research Society* 25: 39-58.

Couper, M. P., and F. Kreuter. 2013. "Using Paradata to Explore Item Level Response Times in Surveys." *Journal of the Royal Statistical Society Series a-Statistics in Society* 176: 271-86.

Couper, Mick P., and Frauke Kreuter. 2013. "Using Paradata to Explore Item Level Response Times in Surveys." *Journal of the Royal Statistical Society* 176: 271-86.

Davis, Rachel E., Mick P. Couper, Nancy K. Janz, Cleopatra H. Caldwell, and Ken Resnicow. 2010. "Interviewer Effects in Public Health Surveys." *Health Education Research* 25: 14-26.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley & Sons.

ESS7. 2015. "Ess7 - 2014 Documentation Report." Retrieved February 22, 2016. http://www.europeansocialsurvey.org/docs/round7/survey/ESS7_data_documentation_report_e01_1.pdf.

Fowler, Floyd J. 1991. "Reducing Interviewer-Related Error through Interviewer Training, Supervision, and Other Means." In *Measurement Errors in Surveys*, edited by Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz and Seymour Sudman, 259-78. Hoboken, NJ: John Wiley & Sons.

Fowler, Floyd J., and Thomas W. Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park, CA: SAGE Publications.

Freeman, John, and Edgar W. Butler. 1976. "Some Sources of Interviewer Variance in Surveys." *Public Opinion Quarterly* 40: 79-91.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2 ed. Hoboken, NJ: John Wiley & Sons.

Groves, Robert M., and Lou J. Magilavy. 1986. "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys." *Public Opinion Quarterly* 50: 251-66.

Hansen, Morris H., William N. Hurwitz, and Max A. Bershad. 1961. "Measurement Errors in Censuses and Surveys." *Bulletin of the International Statistical Institute* 38: 359-74.

Heerwegh, Dirk. 2003. "Explaining Response Latencies and Changing Answers Using Client-Side Paradata from Web Survey." *Social Science Computer Review* 21: 360-73.

Hox, Joop J. 1994. "Hierarchical Regression Models for Interviewer and Respondent Effects." *Sociological Methods & Research* 22: 300-18.

Jenkins, Cleo R., and Don A. Dillman. 1997. "Towards a Theory of Self-Administered Questionnaire Design." In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith De Leeuw, Cathryn Dippo, Norbert Schwarz and Dennis Trewin, 165-96. Hoboken, NJ: John Wiley & Sons.

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3: 305-60.

Keele, Luke John. 2008. *Semiparametric Regression for the Social Sciences*. Hoboken, NJ: John Wiley & Sons.

Kish, Leslie. 1962. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association* 57: 92-115.

Kneip, Thorsten, Frederic Malter, and Gregor Sand. 2015. "Fieldwork Monitoring and Survey Participation in Fifth Wave of SHARE." In *SHARE Wave 5: Innovations & Methodology*, edited by Frederic Malter and Axel Börsch-Supan, 102-59. Munich: MEA, Max Planck Institute for Social Law and Social Policy.

Kosyakova, Yuliya, Jan Skopek, and Stephanie Eckman. 2015. "Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach." *International Journal of Public Opinion Research* 27: 417-31.

Kreuter, Frauke. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: John Wiley & Sons.

Lipps, Oliver, and Grant Benson. 2005. "Cross National Contact Strategies." Paper presented at the AAPOR - ASA Section on Survey Research Methods. Miami Beach, FL.

Lipps, Oliver, and Alexandre Pollien. 2011. "Effects of Interviewer Experience on Components of Nonresponse in the European Social Survey." *Field Methods* 23: 156-72.

Mangione, Thomas W., Floyd J. Fowler, and Thomas A. Louis. 1992. "Question Characteristics and Interviewer Effects." *Journal of Official Statistics* 8: 293-307.

Matschinger, Herbert, Sebastian Bernert, and Matthias C. Angermeyer. 2005. "An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview." *Journal of Official Statistics* 21: 657-74.

Mulligan, Kenneth, J. Tobin Grant, Stephen T. Mockabee, and Joseph Quin Monson. 2003. "Response Latency Methodology for Survey Research: Measurement and Modeling Strategies." *Political Analysis* 11: 289-301.

Olson, Kristen, and Ipek Bilgen. 2011. "The Role of Interviewer Experience on Acquiescence." *Public Opinion Quarterly* 75: 99-114.

Olson, Kristen, and Andy Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71: 273-86.

Olson, Kristen, and Jolene D. Smyth. 2015. "The Effect of Cati Questions, Respondents, and Interviewers on Response Time." *Journal of Survey Statistics and Methodology* 3: 361-96.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63: 581-92.

Sakshaug, Joseph W. 2013. "Using Paradata to Study Response to within-Survey Requests." In *Improving Surveys with Paradata*, edited by Frauke Kreuter, 171-90. Hoboken, NJ: John Wiley & Sons.

Schnell, Rainer, and Frauke Kreuter. 2005. "Separating Interviewer and Sampling-Point Effects." *Journal of Official Statistics* 21: 389-410.

West, Brady T., Frauke Kreuter, and Ursula Jaenichen. 2013. "“Interviewer” Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?". *Journal of Official Statistics* 29: 277-97.

Wooldridge, Jeffrey. 2013. *Introductory Econometrics: A Modern Approach*. 5 ed. Mason, OH: South-Western Cengage Learning.

Yan, Ting, and Kristen Olson. 2013. "Analyzing Paradata to Investigate Measurement Error." In *Improving Surveys with Paradata. Analytic Uses of Process Information*, edited by Frauke Kreuter, 73-95. Hoboken, NJ: John Wiley & Sons.

Appendix

Appendix 1: Wording of selected introduction texts from SHARE wave 5

Overall SHARE Interview (DN001)	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. *1. Continue*
Record linkage (LI004)	We are now changing the topic. The researchers of this study are interested in people's employment history. Important research questions could be answered with data collected by the [German Pension Fund]. We would like to link interview responses with data of the [German Pension Fund]. For reasons of data protection, this cannot be done without your consent. Giving us your permission is completely voluntary. I would kindly ask your consent to do this. Please take a few minutes and to read this form. IWER: Take the 2 consent forms and hand out 1 to the respondent. *1. Continue*
Local area (HH021)	Please look at card [SHOWCARD_ID]. I am now going to read some statements concerning the way you may feel about your local area, that is everywhere within a 20 minute walk or a kilometre of your home. Please tell me whether you strongly agree, agree, disagree or strongly disagree with each statement. *1. Continue*
Health care (HC001)	Now we have some questions about your health care in the last 12 months, that is your doctor visits, hospital stays, or the medication you took. It is also important to us to learn about how much you paid for it out of pocket. By out of pocket payments we mean payments you made directly to your doctor, hospital or pharmacist without getting reimbursed by your health insurance/national health system/third party payer. This also includes co-payments and deductibles for services partly paid by the health insurance/national health system/third party payer. *1. Continue*
Recall test (CF007)	Now, I am going to read a list of words from my computer screen. We have purposely made the list long so it will be difficult for anyone to recall all the words. Most people recall just a few. Please listen carefully, as the set of words cannot be repeated. When I have finished, I will ask you to recall aloud as many of the words as you can, in any order. Is this clear? IWER: Have booklet ready *1. Continue*
Chair stand (CS001)	The next test measures the strength and endurance in your legs. I would like you to fold your arms across your chest and sit so that your feet are on the floor; then stand up keeping your arms folded across your chest. Like this... IWER: Demonstrate. Start of a @BNon-proxy section@B. No proxy allowed. If the respondent is not capable of answering any of these questions on her/his own, press @BCTRL-K@B at each question. *1. Continue*

Appendix 2: Wording of selected survey outcomes from SHARE wave 5

Refusal to income	**HH017_TotAvHHincMonth** How much was the overall income, after taxes and contributions, that your entire household had in an average month in 2012? 0.99..999999999999999.99
Consent given	**LI003_Consent** *IWER: Did R consent to the record linkage?* *Assist respondent if necessary. Hand out R prepared consent form. Assist R if necessary. Cross the form if R refuses. Please insert the consent form in the envelope [addressed DRV] and bring it to the mail box.* 1. Yes. Respondent consented, completed the form and returned the form to me in the envelope. 2. Yes. Respondent consented, but will complete the form later and sent it back himself/herself. 5. No, respondent did not consent to record linkage
Payed anything out of pocket	**HC082_OOPDocsYesNo** Did you pay anything out of pocket for your doctor visits [past your deductible] (in the last twelve months)? Please also include expenses for diagnostic exams, such as imaging or laboratory diagnostics. 1. Yes 5. No
Feeling part	**HH022_LocalFeelPart** I really feel part of this area. Would you say you strongly agree, agree, disagree or strongly disagree? IWER: Show card (SHOWCARD_ID) 1. Strongly agree 2. Agree 3. Disagree 4. Strongly disagree
Cleanliness	**HH024_LocalClean** This area is kept very clean. (Would you say you strongly agree, agree, disagree or strongly disagree?)
Help available	**HH025_LocalPeopleHelpful** If I were in trouble, there are people in this area who would help me. (Would you say you strongly agree, agree, disagree or strongly disagree?)
Vandalism	**HH023_LocalVandalism** Vandalism [m]or[/m] crime [m]is[/m] a big problem in this area. (Would you say you strongly agree, agree, disagree or strongly disagree?)
Amount of words	**CF104_Learn1** Now please tell me all the words you can recall. 1. Hotel 5. Gold 9. King 2. River 6. Market 10. Book 3. Tree 7. Paper 96. None of these 4. Skin 8. Child Interviewer enters the words the respondent correctly recalls. 4 sets of wording lists are assigned randomly (CF104_Learn1-CF107_Learn1).
Compliance with chair stand test	**CS002_Safe** Do you think it would be safe for you to try to stand up from a chair without using your arms? 1. Yes 5. No

Appendix 3: Description of variables

Appendix 3.1: Survey outcomes

Variable	Mean	SD	Min	Max	Description
Refusal income	0.07	0.25	0	1	Refusal to income question
Consent to record linkage	0.66	0.47	0	1	Consent to record linkage
Payed anything out-of-pocket	0.45	0.50	0	1	Payed anything out of pocket for doctoral visits
Feeling part	0.82	0.23	0	1	Feeling part of this area
Vandalism or crime	0.26	0.28	0	1	Vandalism/crime is a big problem in this area
Cleanliness	0.72	0.24	0	1	Area is kept very clean
Help in area	0.75	0.25	0	1	People in this area would help me, when I am in trouble
Amount of words	5.37	1.79	0	10	Number of words recalled in 10 words learning test (first recall)
Compliance to test	0.86	0.35	0	1	Compliance to chair stand measurement

Data: SHARE Wave 5. Unweighted data.

Appendix 3.2: Control variables

Variable	Mean	SD	Min	Max	Description
Days in field	122.20	77.77	0	307	A running number for the days the country is in the field conducting interviews
Female	0.56	0.50	0	1	Gender of respondent (0=male, 1=female)
Age	66.34	10.10	26	103	Age of respondent
Years of education	11.14	4.39	0	25	Years of education (respondent)
Working Status					Working status
1.retired	0.56	0.50	0	1	
2.(self-)employed	0.29	0.45	0	1	
3.other	0.15	0.36	0	1	
Household income					Total income received by all household members in last month
1. Quartile.	0.20	0.40	0	1	
2. Quartile	0.21	0.40	0	1	
3. Quartile	0.21	0.41	0	1	
4. Quartile	0.21	0.40	0	1	
5. Missing	0.18	0.38	0	1	

Appendix 3.2 continued

Household size	2.14	0.95	1	11	Number of persons living in household
Urban area	0.41	0.49	0	1	Living in a large town, suburbs or outskirts of a big city or a big city (reference is rural=0, small town, rural area, or village)
Self-perceived health	0.47	0.27	0	1	Self-perceived health of respondent (5-point scale from 1=excellent to 5=poor, rescaled to ranging from 0=poor to 1=excellent)
Iadl limitations	0.15	0.36	0	1	Having one or more limitations with instrumental activities of daily living (iadl)
Grip strength	32.55	13.37	0	90	Maximum grip strength measurement of respondent
Social activities	0.88	0.33	0	1	Participation in any social activity during the last year (voluntary work, sports, politically active, educational course, playing cards, games)
willing to respond	0.92	0.28	0	1	Willingness to respond to interview was coded by interviewer with very good on a 4-point-scale with the categories bad, fair, good or very good.
Times participated	2.24	1.38	1	5	Number of times the respondent participated in SHARE
Interrupted pattern	0.09	0.28	0	1	Interrupted participation pattern (reference is constant participation)
Longitudinal interview	0.63	0.48	0	1	Interview version (reference: baseline interview)
Late in sample	0.10	0.29	0	1	Within the last 10% of interviews in sample
Second respondent	0.31	0.46	0	1	Interview of second respondent in household
Interviews per day	1.80	1.13	1	13	Number of interviews started per interviewer within one day

Data: SHARE Wave 5. Unweighted data.

Appendix 4: Full regression table for change in reading time

	Overall SHARE Interview	Record linkage	Health care expenditures	Local area information	Recall test	Chair stand
NUMBER OF INTERVIEWS						
1st to 2nd	-2.036***	-7.358	-2.645***	-2.018***	-4.110***	-2.616***
interview	(0.485)	(5.863)	(0.266)	(0.192)	(0.491)	(0.587)
2nd to 10th	-0.302***	-1.711*	-0.487***	-0.348***	-0.735***	-0.737***
interview	(0.072)	(0.804)	(0.041)	(0.032)	(0.072)	(0.093)
10th to 50th	-0.058***	-0.549	-0.073***	-0.062***	-0.099***	-0.159***
interview	(0.014)	(0.309)	(0.011)	(0.009)	(0.016)	(0.025)
50th to last	-0.015	-0.247	-0.013	-0.001	-0.008	-0.004
interview	(0.011)	(0.219)	(0.010)	(0.005)	(0.011)	(0.019)
CONTROLS						
Days in field	0.008	0.050	0.006*	0.002	0.016**	0.007
	(0.004)	(0.082)	(0.003)	(0.003)	(0.005)	(0.007)
Female	1.182***	6.344	0.515**	0.400*	1.082**	1.118*
	(0.348)	(3.425)	(0.189)	(0.159)	(0.334)	(0.441)
Age	-0.001	-0.239	-0.025*	0.003	0.045*	0.019
	(0.018)	(0.200)	(0.010)	(0.009)	(0.018)	(0.023)
Years of	-0.003	0.102	0.057**	0.037*	0.057	0.088*
education	(0.033)	(0.405)	(0.019)	(0.016)	(0.036)	(0.044)
Working status *(reference: other)*						
Retired	-0.124	-6.292	0.523*	0.141	0.134	0.349
	(0.375)	(4.543)	(0.218)	(0.179)	(0.379)	(0.487)
(self-)employed	-0.343	-7.072	0.202	-0.157	-0.226	0.016
	(0.399)	(3.991)	(0.223)	(0.192)	(0.391)	(0.509)
Household income *(ref.2nd quartile)*						
1st quartile	-0.507	-0.277	0.101	-0.105	-0.262	0.014
	(0.380)	(5.966)	(0.260)	(0.193)	(0.458)	(0.528)
3rd quartile	-0.257	-1.773	0.430*	0.394*	0.063	0.094
	(0.459)	(3.380)	(0.205)	(0.175)	(0.393)	(0.509)
4th quartile	-0.783	-3.180	0.163	-0.211	-0.874	-0.556
	(0.416)	(3.802)	(0.228)	(0.182)	(0.461)	(0.549)
Income Missing	0.481	-10.172**	-0.211	-0.390*	-0.549	-1.008
	(0.396)	(3.082)	(0.245)	(0.192)	(0.434)	(0.524)

Appendix 4 continued

Household size	-1.242***	-1.922	-0.245**	0.177**	-0.085	-0.051
	(0.148)	(1.459)	(0.089)	(0.063)	(0.148)	(0.167)
Urban area	0.260	-9.142*	-0.322	-0.001	0.395	-0.425
	(0.315)	(4.073)	(0.217)	(0.162)	(0.364)	(0.456)
Self-perceived health	0.670	-6.768	-0.391	-0.250	-0.493	1.019
	(0.566)	(5.966)	(0.279)	(0.240)	(0.552)	(0.692)
Iadl limitations	0.406	-6.657	-0.159	-0.356	0.476	-2.930***
	(0.401)	(3.943)	(0.212)	(0.183)	(0.410)	(0.556)
Grip strength	0.030*	0.211	0.032***	0.014*	0.055***	0.080***
	(0.015)	(0.145)	(0.007)	(0.007)	(0.013)	(0.019)
Social activities	-0.595	1.058	0.930***	0.793***	0.877*	1.692**
	(0.444)	(5.994)	(0.269)	(0.214)	(0.426)	(0.527)
Very good willingness to respond	-0.094	25.530***	0.402	0.736**	1.136*	2.444***
	(0.443)	(5.292)	(0.293)	(0.276)	(0.469)	(0.584)
Times participated	-0.286	0.626	0.016	0.029	-0.476**	0.341
	(0.193)	(7.358)	(0.087)	(0.077)	(0.158)	(0.223)
Interrupted pattern	-1.102	2.505	0.112	-0.156	-0.270	-0.075
	(0.566)	(14.469)	(0.257)	(0.280)	(0.440)	(0.582)
Longitudinal interview	0.620	-1.053	0.065	0.290	-1.071*	-0.028
	(0.600)	(26.942)	(0.273)	(0.261)	(0.518)	(0.667)
Late in sample	0.042	-5.823	0.220	-0.160	-1.515**	0.845
	(0.482)	(5.320)	(0.291)	(0.273)	(0.530)	(0.751)
Second respondent	6.783***	-41.729***	-3.375***	14.638***	-7.052***	-7.850***
	(0.393)	(2.762)	(0.166)	(0.407)	(0.292)	(0.379)
Interviews per day	-0.752***	-0.008	-0.479***	-0.084	-1.054***	-0.635***
	(0.116)	(1.772)	(0.068)	(0.046)	(0.114)	(0.149)
Log-likelihood	-134675.7	-25302.4	-189199.9	-116596.9	-228603.5	-232137.8
R^2 (within)	0.043	0.116	0.084	0.075	0.073	0.048
N $_{\text{Interviews/Respondents}}$	32743	4542	50843	34766	52690	50219
N $_{\text{Interviewers}}$	1565	176	1576	1565	1577	1576

Data: SHARE Wave 5. Panel-robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix 5: Robustness checks on change in reading behavior

Appendix 5.1: Logarithmic transformation

	Overall SHARE Interview	Record linkage	Health care expenditures	Local area information	Recall test	Chair stand
NUMBER OF INTERVIEWS						
1st to 2nd	-0.171***	-0.024	-0.192***	-0.161***	-0.150***	-0.184***
interview	(0.027)	(0.089)	(0.017)	(0.015)	(0.017)	(0.024)
2nd to 10th	-0.043***	-0.027*	-0.041***	-0.034***	-0.035***	-0.051***
interview	(0.004)	(0.012)	(0.003)	(0.003)	(0.003)	(0.004)
10th to 50th	-0.009***	-0.009	-0.010***	-0.010***	-0.007***	-0.013***
interview	(0.001)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)
50th to last	-0.001	-0.009	-0.001	-0.001	-0.001	-0.001
interview	(0.001)	(0.006)	(0.001)	(0.001)	(0.001)	(0.001)
Log-likelihood	-42830.1	-6592.0	-56188.8	-33642.6	-68441.7	-75690.1
R^2 (within)	0.071	0.169	0.117	0.086	0.077	0.073
$N_{\text{Interviews/Respondents}}$	32743	4542	50843	34766	52690	50219
$N_{\text{Interviewers}}$	1565	176	1576	1565	1577	1576

Data: SHARE Wave 5. Fixed-effects regressions. Panel-robust standard errors in parentheses.; Controlled for days in field, respondent characteristics, and interview specifics. Weighted results. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix 5.2: Exclusion of interviewers with extreme total number of interviews

Here, we excluded interviewers who conducted very few or very many interviews. We set the thresholds at the 5th and 95th percentile, which resulted in dropping interviewers who conducted less than 15 or more than 148 interviews in total.

	Overall SHARE Interview	Record linkage	Health care expenditures	Local area information	Recall test	Chair stand
NUMBER OF INTERVIEWS						
1st to 2nd	-2.239***	-4.312	-2.454***	-1.748***	-4.193***	-1.995***
interview	(0.411)	(5.030)	(0.252)	(0.178)	(0.437)	(0.577)
2nd to 10th	-0.298***	-1.799*	-0.503***	-0.348***	-0.727***	-0.779***
interview	(0.062)	(0.717)	(0.041)	(0.031)	(0.068)	(0.093)
10th to 50th	-0.047***	-0.750**	-0.072***	-0.059***	-0.085***	-0.155***
interview	(0.014)	(0.286)	(0.012)	(0.008)	(0.016)	(0.025)
50th to last	-0.014	-0.305	-0.017	-0.000	-0.010	-0.005
interview	(0.016)	(0.223)	(0.013)	(0.007)	(0.015)	(0.025)
Log-likelihood	-123960.1	-24294.2	-173533.1	-106671.1	-208546.5	-213586.4
R^2 (within)	0.046	0.112	0.079	0.072	0.072	0.046
$N_{\text{Interviews/Respondents}}$	30116	4375	46472	31720	48002	45941
$N_{\text{Interviewers}}$	1241	155	1241	1239	1241	1241

Data: SHARE Wave 5. Fixed-effects regressions. Panel-robust standard errors in parentheses.; Controlled for days in field, respondent characteristics, and interview specifics. Weighted results. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix 5.3: Learning effects in balanced panels

Ideally, we would like to create subgroups of interviewers that have conducted exactly the same total number of interviews. Unfortunately, this is not feasible with our data, because subgroups then become too small for reliable estimates. To avoid very low numbers of interviews within one subset of interviewers, we therefore first grouped interviewers in intervals that have a sufficient number of interviews to obtain reliable coefficients. We then restricted the analyses to the lowest total number of interviews within each subset to construct balanced panels. This results in group sizes between 204 and 304 interviewers. The number of interviewers per group is lower for record linkage ranging between 8 and 43 interviewers, because here we only have data from Germany. The results for the learning effect (spline between 2nd and 10th interview) are displayed below.

		Total number of interviews					
Intro item	**All interviews**	**10-19**	**20-29**	**30-39**	**40-49**	**50-69**	**70+**
Intro to overall SHARE interview	-0.302***	-0.560	-0.459**	-0.230	-0.074	-0.411**	-0.092
	(0.072)	(0.289)	(0.148)	(0.173)	(0.152)	(0.130)	(0.155)
Intro to record linkage	-1.711*	-5.404	-2.438	-1.216	-0.853	-	-
	(0.804)	(3.158)	(1.448)	(1.308)	(1.617)		
Intro to health care expenditures	-0.487***	-0.400**	-0.579***	-0.337***	-0.486***	-0.585***	-0.312***
	(0.041)	(0.141)	(0.094)	(0.099)	(0.097)	(0.086)	(0.071)
Intro to local area information	-0.348***	-0.494***	-0.388***	-0.213**	-0.356***	-0.448***	-0.375***
	(0.032)	(0.119)	(0.089)	(0.080)	(0.076)	(0.069)	(0.056)
Intro to recall test	-0.735***	-0.969***	-0.719***	-0.766***	-0.528**	-0.692***	-0.494***
	(0.072)	(0.258)	(0.176)	(0.194)	(0.159)	(0.124)	(0.113)
Intro to chair stand	-0.737***	-0.169	-0.905***	-0.832***	-0.454	-0.733***	-0.874***
	(0.093)	(0.313)	(0.244)	(0.223)	(0.252)	(0.200)	(0.243)

Note: Cell entries are spline coefficients indicating the 2nd to 10th interview and respective standard errors. Only coefficients based on more than ten interviewers conducting at least ten interviews are displayed. Panel-robust standard errors in parentheses.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix 6: Full regression table for survey outcomes

Appendix 6.1: Change in survey outcomes – first group

	Refusal income	Consent to record linkage	Payed for doctoral visits	Recall 10 words list – first trial	Compliance chair stand
Duration of respective intro item	0.000	0.001***	0.000	0.001*	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Days in field	0.000	-0.001**	-0.000	0.000	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	0.007	0.017	0.014*	0.616***	0.102***
	(0.006)	(0.020)	(0.007)	(0.028)	(0.006)
Age	0.000	-0.003*	-0.001	-0.045***	-0.003***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)
Years of education	0.002***	0.003	0.004***	0.077***	0.001
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
Working status *(reference: other)*					
Retired	0.008	0.012	0.017*	0.160***	0.042***
	(0.007)	(0.027)	(0.008)	(0.029)	(0.007)
(self-)employed	0.019**	-0.059*	0.016	0.071*	0.010
	(0.007)	(0.025)	(0.008)	(0.029)	(0.006)
Household income *(ref.2nd quartile)*					
1st quartile		-0.108**	-0.010	-0.095**	-0.003
		(0.038)	(0.010)	(0.034)	(0.007)
3rd quartile		0.015	0.023**	0.117***	0.003
		(0.023)	(0.008)	(0.027)	(0.006)
4th quartile		-0.023	0.028**	0.169***	-0.012
		(0.031)	(0.009)	(0.030)	(0.006)
Income Missing		-0.192***	0.005	0.037	-0.013
		(0.029)	(0.009)	(0.030)	(0.007)
Grip strength	-0.000	0.001	-0.000	0.016***	0.007***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)
Social activities	0.013	0.074	0.030**	0.518***	0.055***
	(0.010)	(0.040)	(0.009)	(0.035)	(0.008)
Willing to respond	-0.096***	0.191***	0.022*	0.621***	0.056***
	(0.012)	(0.036)	(0.010)	(0.042)	(0.010)
Times participated	-0.004	-0.055	0.004	0.026*	0.005*
	(0.002)	(0.050)	(0.003)	(0.011)	(0.002)

Appendix 6.1 continued

Interrupted pattern	0.002 (0.006)	0.073 (0.088)	0.005 (0.010)	-0.068* (0.034)	-0.006 (0.007)
Longitudinal interview	0.003 (0.007)	0.122 (0.177)	-0.017 (0.011)	0.075* (0.034)	-0.009 (0.008)
Late in sample	0.007 (0.010)	0.013 (0.038)	-0.036** (0.012)	-0.029 (0.038)	0.006 (0.008)
Second respondent	-0.008* (0.004)	0.063*** (0.017)	-0.003 (0.006)	-0.013 (0.020)	0.021*** (0.004)
Interviews per day	0.003 (0.002)	-0.019 (0.010)	0.000 (0.003)	-0.022* (0.009)	-0.002 (0.002)
Log-likelihood	1595.1	-2512.1	-17481.1	-90700.4	-7381.5
R^2 (within)	0.012	0.079	0.006	0.265	0.250
N $_{\text{Interviews/Respondents}}$	28268	4542	44937	52203	49984
N $_{\text{Interviewers}}$	1547	176	1574	1576	1576

Data: SHARE Wave 5. Panel-robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix 6.2: Change in survey outcomes – second group (local area variables)

	Feeling part	Cleanliness	Help available	Vandalism or crime
Duration of respective intro item	-0.000 (0.000)	-0.001 (0.000)	0.001 (0.000)	-0.001* (0.000)
Days in field	-0.000 (0.000)	0.000** (0.000)	0.000 (0.000)	-0.000 (0.000)
Female	-0.000 (0.006)	-0.025*** (0.007)	0.025** (0.008)	-0.015 (0.008)
Age	0.002*** (0.000)	0.003*** (0.000)	0.002*** (0.000)	-0.002*** (0.000)
Years of education	-0.002*** (0.001)	0.000 (0.001)	0.002* (0.001)	-0.002** (0.001)
Working status *(reference: other)*				
Retired	-0.006 (0.007)	0.005 (0.009)	0.004 (0.009)	-0.011 (0.009)
(self-)employed	0.002 (0.007)	0.027** (0.010)	0.010 (0.009)	-0.031*** (0.009)

Appendix 6.2 continued

Household income *(ref.2nd quartile)*				
1st quartile	0.000 (0.008)	-0.004 (0.011)	0.007 (0.010)	0.003 (0.011)
3rd quartile	0.004 (0.007)	0.019* (0.008)	0.014 (0.008)	0.004 (0.008)
4th quartile	0.009 (0.007)	0.022* (0.009)	0.027** (0.008)	-0.001 (0.009)
Income Missing	-0.002 (0.007)	0.012 (0.010)	0.018* (0.008)	0.017 (0.010)
Household size	0.010*** (0.002)	0.003 (0.003)	0.002 (0.003)	-0.009** (0.003)
Urban area	-0.013* (0.006)	-0.051*** (0.008)	-0.046*** (0.007)	0.062*** (0.008)
Self-perceived health	0.075*** (0.009)	0.065*** (0.011)	0.100*** (0.011)	-0.068*** (0.012)
Iadl limitations	-0.022*** (0.007)	-0.001 (0.009)	0.003 (0.008)	0.002 (0.009)
Grip strength	0.000 (0.000)	-0.001 (0.000)	0.001* (0.000)	-0.001* (0.000)
Social activities	0.021* (0.009)	-0.002 (0.011)	0.044*** (0.010)	-0.008 (0.011)
Willing to respond	0.027** (0.008)	0.016 (0.011)	0.030* (0.012)	0.007 (0.011)
Times participated	0.003 (0.003)	0.001 (0.003)	-0.004 (0.003)	-0.000 (0.004)
Interrupted pattern	0.005 (0.009)	0.004 (0.011)	-0.012 (0.011)	-0.017 (0.012)
Longitudinal interview	-0.002 (0.010)	-0.014 (0.013)	0.024* (0.011)	-0.001 (0.012)
Late in sample	-0.004 (0.010)	-0.017 (0.011)	-0.001 (0.012)	0.016 (0.013)
Interviews per day	-0.002 (0.002)	-0.001 (0.002)	-0.001 (0.002)	0.002 (0.003)
Log-likelihood	-1232.3	-9372.8	-8164.8	-11729.6
R^2 (within)	0.013	0.011	0.016	0.011
$N_{\text{Interviews/Respondents}}$	34664	34692	34168	34607
$N_{\text{Interviewers}}$	1565	1565	1563	1565

Data: SHARE Wave 5. Panel-robust standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix 7: Paying Out of Pocket by Average OOP-Level in a Country

	Countries with avg. OOP-level under 25%	Countries with avg. OOP-level between 25 and 75%	Countries with avg. OOP-level over 75%
Reading duration of intro to health care	-0.04	-0.04*	0.04
Constant	-0.13*	0.53***	0.93***
N $_{\text{Interviews/Respondents}}$	12479	23996	8462
N $_{\text{interviewers}}$	389	885	300

Data: SHARE Wave 5. Fixed-effects regressions. Panel-robust standard errors in parentheses; Controlled for days in field, respondent characteristics, and interview specifics. Weighted results. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
OOP=out of pocket payment.