READING FAST, READING SLOW: THE EFFECT OF INTERVIEWERS' SPEED IN READING INTRODUCTORY TEXTS ON RESPONSE BEHAVIOR

MICHAEL BERGMANN
JOHANNA BRISTLE*

Guidelines for interviewers frequently include instructions to read question texts exactly as they are worded. Deviations from these guidelines on standardized interviewing might affect the comparability of survey answers and impair the quality of data. This paper contributes to the literature on interviewer behavior by analyzing how interviewers change their reading behavior during fieldwork and whether this behavioral change influences the response behavior of survey respondents. We use item-level paradata from the Survey of Health, Ageing and Retirement in Europe (SHARE) to measure interviewers' reading times and focus our analyses on introductory questions that do not require an immediate response by the respondent. In contrast to prior research, this focus enables us to disentangle the reading times of interviewers from the response times of respondents. Based on fixed effects regression, our results show systematic changes in interviewers' reading times of

MICHAEL BERGMANN and JOHANNA BRISTLE are with the Technical University of Munich (Chair for the Economics of Aging), Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Amalienstr. 33, 80799 Munich, Germany

The authors are grateful for valuable discussions with colleagues at several conferences and MEA, especially Julie Korbacher, Raphael Guber, Helmut Farbmacher and Annette Scherpenzeel. This article uses data from the Survey of Health, Ageing, and Retirement in Europe (SHARE). The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: N°211909, SHARE-LEAP: N°227822, SHARE M4: N°261982). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, 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, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

*Address correspondence to Johanna Bristle, Max Planck Institute for Social Law and Social Policy, Amalienstr. 33, 80799 Munich, Germany; E-mail: bristle@mea.mpiisoc.mpg.de.

introductory items: First, reading times significantly decrease over the survey's field period, even after controlling for relevant respondent characteristics and specific aspects of the interview situation. Second, a cross-national comparison, including fourteen European countries plus Israel, reveals that the decrease is uniform in almost all countries, suggesting its generalizability across different cultural contexts. Third, the decrease in reading times influences response behavior to varying degrees. Response behavior is affected if introductions contain relevant information for understanding or fulfilling the required task and especially if the response refers to within-survey requests. On the basis of these findings, we discuss the possible consequences for questionnaire design, interviewer training, and fieldwork monitoring.

KEYWORDS: Cross-national survey; Interviewer behavior; Measurement error; Paradata; Reading time; Response behavior; SHARE.

1. INTRODUCTION

In the total survey error framework, the goal for obtaining accurate survey estimates is defined as minimizing error in the several quality components that contribute to the total error of an individual survey estimate (Groves and Lyberg 2010; Lyberg 2012). While bias can occur at any step throughout the process of designing, collecting, and calculating the survey estimate, the role of the interviewer is of particular interest within this framework (West and Blom 2017). In terms of measurement error during face-to-face interviews, standardized interviewing is commonly seen as a way to guarantee that each respondent is exposed to the same experience during questioning to achieve comparable answers (e.g., Fowler and Mangione 1990; Fowler 1991). According to Groves et al. (2009, p. 305), “[u]niform wording of questions asked is perhaps the most fundamental and universally supported principle of standardized interviewing.” Deviations from standardized interviewing might produce inconsistent measurements and impair data quality. Therefore, guidelines for interviewers frequently include the instruction to read question texts exactly as they are worded (Malter 2013).

Previous research, however, has long shown that interviewers do not always follow the rules of standardized interviewing (see Hyman and Cobb 1954 for an early example). Instead, interviewers adapt their interviewing behavior as they gain more experience, leading to declining interview times over the field period (Olson and Peytchev 2007; Böhme and Stöhr 2014). This decline has been shown to be more pronounced in face-to-face settings than in telephone interviews, presumably due to less strict monitoring (Kirchner and Olson 2017). Although there is evidence that interviewers do not adhere strictly to survey instructions as their experience with the procedures increases (Chromy,

Eyerman, Odom, McNeeley, Hughes 2005; Olson and Bilgen 2011), the question of how this might substantially influence the response behavior of respondents is less clear. In an early work based on telephone interviews, Groves and Magilavy (1986) found significant variation in interviewers' reading behavior but a rather low correlation with survey answers. Cannell, Marquis, and Laurent (1977) found that longer question times increase the amount of information provided and the accuracy of the responses. Additionally, Holbrook et al. (2016) recently showed that the reading difficulties of interviewers were associated (although not significantly) with the comprehension problems of respondents, in turn possibly influencing their response behavior. Other studies considering interviewer effects (Kish 1962; Mangione, Fowler, and Louis 1992; Davis, Couper, Janz, Caldwell, and Resnicow 2010; West, Kreuter, and Jaenichen 2013; Josten and Trappmann 2016; West, Conrad, Kreuter, and Mittereder 2018) only marginally probe their reading behavior and are often limited to a single country.

Therefore, the central aim of this article is to shed light on the question of whether (and how) changes in interviewers' reading times—and thus a deviation from standardized interviewing—influence response behavior in a cross-national face-to-face survey. In a cross-national setting involving different cultural contexts, comparability and standardization of interviewing are of utmost importance because otherwise respondents' answers cannot be meaningfully analyzed (Loosveldt and Beullens 2013).

From a theoretical point of view and based on the total survey error framework, several reasons can lead to an interviewer's deviating reading behavior that in turn might influence measurement error. Regarding the survey instrument, difficult, sensitive, or poorly designed questions can motivate interviewers into making adjustments to assist respondents (Mangione et al. 1992; Schnell and Kreuter 2005). Furthermore, interviewers might be more or less inclined to adopt conversational practices toward respondents to build rapport (Schaeffer, Dykema, and Maynard 2010) or to assure that respondents interpret questions correctly (Schober and Conrad 1997). Similarly, Ongena and Dijkstra (2007) argue that instead of reading questions as worded, interviewers, as their experience increases, recall questions from memory. In addition, interviewers might paraphrase questions depending on question complexity and politeness norms or omit lengthy qualifying statements from questions (Ongena 2003; Ongena and Dijkstra 2007). Such rewording might be used to tailor questions to respondents' needs and can include skipping specific questions (Ackermann-Piek and Massing 2014; Kosyakova, Skopek, and Eckman 2015), shortening items (Haan, Ongena, and Huiskes 2013), or speeding through sections of the questionnaire (Fowler and Mangione 1990; Matschinger, Bernert, and Angermeyer 2005; Brüderl, Huyer-May, and Schmiedeberg 2012; Haan et al. 2013). We argue that the latter becomes especially relevant when interviewers are paid per completed interview (rather than per hour) because then

interviewers have a strong monetary incentive to conduct their assigned interviews quickly.

Based on these considerations, we formulate three hypotheses. First, we expect that only a small proportion of interviewers will strictly follow the rules of standardized interviewing (i.e., read introductory texts exactly as they are worded) and that most will adjust their reading behavior over the survey's field period. These adjustments, due to increasing experience with the survey procedure, will result in a decrease in reading time, even when controlling for respondents' characteristics and specific aspects of the interview situation. Second, following the argument that it is rational to shorten interviews when payment is based on completed interviews, as is the case for all countries in our analyses, we expect that a decrease in reading time is uniform across countries and can thus be generalized. Third, we further hypothesize that the amount of informational content that is shortened or read out very quickly will affect response behavior to a varying extent. The largest effects are expected with respect to within-survey requests, which ask respondents to agree to the collection of additional data and go beyond the traditional question-answer sequence (Sakshaug 2013).

We base our analyses on data from a large-scale face-to-face survey of fourteen European countries plus Israel. Because the reading and response times of interviewers and respondents are often intertwined, we focus on introductory items that do not require a response by the respondent. As these items rarely involve interactions between interviewers and respondents, they are a much cleaner measure in our view than previously used indicators such as interviewing speed or interview length. The results are reported in two steps: first, we investigate how reading times of introductory items change over the survey's field period, taking into account country differences. Second, we explore whether these changes in reading behavior affect collected survey responses. We conclude with a discussion of the implications of our findings with respect to questionnaire design, interviewer training, and fieldwork monitoring.

2. DATA AND METHODS

2.1 Data

We use data from Wave 5 of the Survey of Health, Ageing and Retirement in Europe (SHARE, Börsch-Supan 2015), a multidisciplinary, cross-national panel study that explores the aging process of individuals aged fifty and older and their partners (see Börsch-Supan, Brandt, Hunkler, Kneip, Korbmacher, et al. 2013) for methodological details. SHARE is conducted biannually. The fifth wave was collected in 2013 in fourteen European countries (Austria, Belgium, the Czech Republic, Denmark, Estonia, France, Germany, Italy, the Netherlands, Luxembourg, Slovenia, Spain, Sweden, and Switzerland) and

Israel. Data collection was conducted face-to-face using a centrally developed Computer Assisted Personal Interview (CAPI) system, and all interviewers were instructed to follow the standardized interviewing approach and to read question texts as scripted (Malter 2013). The samples in all countries were based on probability sampling with full coverage of the 50+ population (for details on the national sampling frames see De Luca, Rossetti, and Malter 2015). The household response rate for refreshment samples in Wave 5 varied from 33 percent to 61 percent across countries, according to standards set by the American Association for Public Opinion Research (see AAPOR 2016, RR3). Retention rates for individuals who participated in the previous wave, including recovery, varied between 73 percent and 89 percent (Bergmann, Kneip, De Luca, and Scherpenzeel 2017). Magnitude and differences in country were comparable to other cross-national surveys in Europe (e.g., ESS7 2016).

In addition to the released CAPI data, we used item-level paradata derived from Blaise audit trails, also called keystroke data (Kreuter 2013). SHARE keystroke data contains timestamps for all actions recorded via keyboard or mouse along the question-answer process. From this data, source latent timers were extracted; they measure the exact duration from the moment when an item appears on the screen until the interviewer confirms the entry of an answer or clicks “continue” in the case of an introductory item to switch to the next question.

Keystroke data is available for all CAPI interviews. However, similar to adjustments utilized in response latency analyses, we first excluded item durations below one and above 1,000 second(s), as they were assumed to be caused by technical errors or interview interruptions (Bristle 2015). Second, outliers were investigated country by country and were excluded from the analyses when they were greater than two standard deviations below or above the respective mean (see, e.g., Couper and Kreuter 2013). Our final sample for analysis consists of 63,397 interviews conducted by 1,581 interviewers in fifteen countries. Thus, on average, every interviewer conducted 40.1 interviews but with some variation across countries (see table A3 in the Appendix).

2.2 Measurement of the Main Variables of Interest

Our main variable of interest is the reading time of introductory items based on latent timers, as described previously. In contrast to classical survey questions, introductory items have the advantage of not requiring an immediate response by the respondent. Therefore, reading times of introductory items measure how fast or slow an interviewer reads out information to the respondent, and we use these durations as a proxy for interviewers’ reading behavior. Thus, we assume that the time an interviewer needs to read out an introductory item can be used to capture the deviation from standardized interviewing over the survey’s field period, despite recognizing that we are not able to distinguish

different mechanisms leading to a specific reading time due to a lack of additional information (e.g., from audio recordings) (Ackermann-Piek and Massing 2014). We selected six introductory items that contain relevant information for respondents to provide a well-informed answer and used only cases that were not backed up, edited, or re-entered. We differentiate between three types of informational content that appear relevant for a correct understanding of the survey question or 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.

Regarding information on respondents' confidentiality concerns, we used an introductory statement about the explicit acceptance of refusing to answer sensitive questions and another about the respondents' consent to link the data to administrative records. The record linkage project is implemented only in Germany, and our analyses in this case are therefore restricted to the German subsample. Introductions containing precise definitions of difficult or unfamiliar terms include an introduction text to health care expenditures, thus explaining the term "out-of-pocket payments," and a second definition introducing questions about the respondent's "local area," which is defined as the geographical region within a twenty-minute walk or one kilometer around the respondent's home. Lastly, concerning items with detailed instructions helping respondents to understand how to conduct a specific task, we used a cognitive recall test, asking respondents to name up to ten words from a previous given list within a certain time, and instructions for the physical measurement of strength and endurance in the respondents' legs (chair stand).

In addition to investigating changes in interviewers' reading duration over time, we analyzed the impact of this change on the respondents' response behavior, which would be affected by the way the six introduction texts were read out by the interviewer, as they contained relevant information for the respondent to provide a valid answer or fulfill a requested task. A brief overview of the introductory items used and the corresponding survey responses is presented in [table 1](#). The exact wording, answer categories, and distributions can be found in the Appendix.

Regarding items about confidentiality concerns, we suppose that not reading this information thoroughly will result in a lower amount of refusals and thus more answers to sensitive questions, such as the respondents' income, because they are not aware that refusing to answer is acceptable. In contrast, we expect more refusals and thus a lower consent rate to record linkage if the interviewers superficially speed through the introduction of the linkage procedure. This outcome is due to the very different framing of the record linkage procedure. Here, regulatory frameworks require the respondent's written consent, which should lead to a much higher awareness. Hence, respondents might feel uncomfortable about their privacy when not being fully informed

Table 1. List of Introductory and Survey Response Items Used

Type of content	Introductory item	Survey response item
Confidentiality	Intro to overall SHARE interview	Refusal to income
Confidentiality	Intro to record linkage	Consent to record linkage
Definition	Intro to health care expenditures	Paid anything out-of-pocket
Definition	Intro local area information	Residential satisfaction
Instruction	Intro to recall test	Amount of words recalled
Instruction	Intro to chair stand	Compliance with chair stand test

about the specific conditions of data linkage. Such an outcome might be especially relevant for older respondents, who are the SHARE target group. For those, trust in the confidentiality of the procedure and the interviewers’ efforts to increase their comprehension are key factors for their consent to link survey data with administrative records (Korbmacher and Schroeder 2013). Regarding items with definitions of uncommon or difficult terms, we expect that omitting or speeding through time-consuming explanations reduces the likelihood of reporting out-of-pocket payments for doctors’ visits. Similarly, we expect that shortening the definition of local area results in less variation in residential satisfaction (i.e., personal affiliation, security, upkeep, and social relations in the neighborhood). Lastly, we expect respondents to perform better after receiving detailed instructions and sufficient time to understand the scripted instructions before starting the test. Good test performance was measured here in terms of more words recalled, indicating the effort respondents put into performing the test thoroughly, and a higher compliance rate in the chair stand test.

2.3 Empirical Model

Our analyses are based on a hierarchical, longitudinal data structure, where interviews are ordered in time and nested in interviewers who themselves are nested in countries. Because we are interested in behavioral changes *within* interviewers over the survey’s field period instead of explaining differences *between* interviewers, we prefer a fixed effects approach over the more commonly applied multilevel modeling techniques. This approach comes with the advantage that we can relax the mandatory assumption of multilevel models (i.e., unobserved variables such as interviewer characteristics or countries’ cultural aspects must be uncorrelated with all observed explanatory variables to achieve unbiased results) (see supplement 2.3 of the online [Supplementary Material](#) for further indications supporting our decision).

Under this assumption, we can write our general model based on the terminology of [Wooldridge \(2013\)](#) as,

$$y_{it} = \beta x_{it} + \alpha_i + u_{it},$$

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

When analyzing changes in interviewers' reading behavior 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 can be rewritten as follows:

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

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, with the fixed effects transformation, we can neglect all stable interviewer characteristics that shape general reading behavior, such as individual base speed of reading, language, cognitive abilities, and the sociodemographic characteristics of the interviewer. The same applies for sampling-point or area effects that are difficult to disentangle from interviewer effects in multilevel models (e.g., [Schnell and Kreuter 2005](#); [Beullens and Loosveldt 2016](#)) and even survey management strategies or cultural aspects of the participating countries. Instead, we need to carefully control for potential time-variant confounders, such as respondent characteristics and specific features of a particular interview, including period effects adjusting for the country-specific progression of fieldwork ("days in field") that might influence the interviewer's reading behavior (see [table A5](#) in the Appendix for a list of all control variables used). Not taking these factors into account runs the risk of attributing shorter reading times to interviewer behavior, although this effect might be confounded with changes in the sample composition. In addition, we weight our data by giving each interviewer an equal importance to avoid interviewers with a very low (high) number of interviews exerting a disproportionately small (large) influence on the results and calculate interviewer-clustered standard errors that are robust to heteroscedasticity and autocorrelation.

Our empirical analysis is organized in two steps. First, we ran linear fixed effects regressions that explain changes in the interviewer's reading behavior as a dependent variable to test our first hypothesis. Moreover, we tested separate models per country to explore if our results are universally applicable (hypothesis 2). In the second step, the change in interviewer's reading duration acted as independent variable x_{it} to explain consequences on the previously

mentioned survey responses, acting as y_{it} (hypothesis 3). For the dichotomous survey responses, we applied logistic fixed effects models. All linear models used the `xtreg, fe` command in Stata 14 with fixed effect estimator and standard errors clustered at the interviewer level. The dichotomous survey responses were estimated using `xtlogit, fe` with centered predictors. All programs are available from the authors upon request.

3. RESULTS

3.1 Change in Interviewers' Reading Behavior Over the Survey's Field Period

After applying the fixed effects approach and controlling for respondent characteristics and the face-to-face interview situation, we obtained a reliable trend of decreasing reading times that is shown in [figure 1](#). Here, predicted reading time is displayed over the order of interviews on the x-axis with 95 percent confidence intervals. We used a transformed log scale to model our assumption of a sharp decrease in interviewers' reading time during the first interviews and a more stable pattern afterwards.

Overall, we see a strong and statistically significant decline for all six introductory items. All coefficients and the respective regression models are reported in online Supplementary Material. As expected, the steepest decrease is observed at the beginning. Here, it can be reasonably assumed that interviewers' reading behavior becomes less standardized as increasing experience leads to deviations from standardized interviewing, such as incorrect reading or skipping (parts) of questions ([Ackermann-Piek and Massing 2014](#)). In addition, external factors (e.g., a better preparation of additional materials), which might play a considerable role between the first and second interview, do not influence our results when applying a parametric log transformation for number of interviews because the reading time for the first interview was excluded from the analyses. After a number of interviews, the decrease is apparently attenuated, as behavioral adjustments largely have taken place.

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 introductory texts displayed previously, except for the record linkage item that was only asked in Germany. Missing values due to questionnaire routing have been imputed by the country-specific mean of the reading time at a particular interview to obtain a comparable measure; otherwise, interviewers who have not read out every selected item, and thus lack those durations by design, would appear much faster than interviewers who read out all items to the respondents. A cumulative measure is used to test if the same interviewers are equally fast with respect to the items we analyzed, thus indicating a general interviewer behavior. Based on this cumulated measure, we again ran fixed effects regressions with the same covariates to control for respondent characteristics and

Figure 1. Change in Reading Behavior Over Order of Interviews. Data: SHARE Wave 5. Fixed effects regressions with log transformation for order of interview. Controlled for days in field, respondent characteristics, and interview specifics. Weighted results.

interview specifics including period effects, but we now distinguish our analysis between countries.

Overall, we again see a comparable strong and statistically significant decline in interviewers' reading time in the beginning that is attenuated afterwards for fourteen out of fifteen countries. This outcome suggests a very similar behavior of interviewers across various countries. What attracts our

Figure 2. Country-Specific Change in Reading Behavior Over Order of Interviews. Data: SHARE Wave 5. Fixed effects regressions with log transformation for order of interview. Controlled for days in field, respondent characteristics, and interview specifics. Weighted results.

attention is the very different (starting) level of interviewers' reading time, which results in a country-specific scaling of the y-axis in the subgraphs of [figure 2](#). We attribute these level differences mostly to linguistic aspects in terms of variation in the amount of characters needed for an expression or the general base speed of a language and to other unobserved country-specific aspects (see [Loosveldt and Beullens 2013](#)). Note that the general pattern of decreasing

reading times is not affected by these differences, as time-constant country specifics were implicitly controlled for in fixed effects models.

There is only one obvious exception from the general pattern of monotonically decreasing reading times; Italy shows declining reading times that stabilize after some interviews. From fieldwork management, we know that the overall interview length and reading times for specific items here were part of the regular monitoring reports that were shared with the interviewers. Therefore, interviewers knew that their behavior was being at least partly observed. Somewhat differently, in Israel, the very low starting level of reading time simply leaves little room for interviewers to further speed up their reading. Additionally, Israel had only very few interviewers in total who, at the same time, conducted many interviews. This leads to larger standard errors and less precise estimates. Hence, the results in this case must be regarded with caution.

3.2 Robustness Checks

We conducted several robustness checks to verify our results (see supplement 2 of the online [Supplementary Material](#)). First, we reran all our models using a logarithmic transformation of the reading time. This process is sometimes recommended in the literature on response times to take into account the typically highly right-skewed distribution of these variables (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 supplement 2.1 of the online [Supplementary Material](#)). Therefore, we kept the more intuitive scaling of reading time in seconds.

The second robustness check addressed the issue that interviewers differ in the total number of their completed interviews. In this respect, one can imagine that the number of interviews conducted by a particular interviewer is not random but rather due to unobserved characteristics such as working part-time, being assigned to this project for only, a short period or having specific skills for gaining participation. This phenomenon would imply a violation of the missing at random assumption (MAR) for unbalanced panels (Rubin 1976) and, hence, limit the generalizability of the fixed effects estimates. Therefore, we created a series of balanced panels by truncating the interview sequences to the same number of interviews within groups (e.g., for all interviewers with twenty to twenty-nine interviews, we used the first twenty interviews to create a balanced panel within this group) and re-estimated our models for these subgroups of interviewers. While effects for reading the introduction to the overall SHARE interview and record linkage are not always statistically significant, the coefficients for all other introductory items are very stable and comparable to the results based on the overall sample (see supplement 2.2 of the online [Supplementary Material](#)).

The third robustness check refers to the use of fixed effects models. In contrast to a random effects model, the fixed effects approach uses only variations within interviewers for the estimation of coefficients. All variations among

interviewers are absorbed by the intercept. In our case, the within variance for all models is larger or only marginally smaller than the between variance. This difference is reflected by intraclass correlation coefficients (ICC) between 23 and 56 percent. By using a fixed effects approach, we are able to implicitly control for this share of variance, which is attributed to time-constant characteristics (e.g., interviewer attributes or country specifics). In addition, the Hausman test as a point of reference does not recommend random effects for any equation (see [Clark and Linzer 2015](#) for a discussion of the test). We report the results of a random-intercept model as a robustness check in supplement 2.3 of the online [Supplementary Material](#). Overall, the results are very similar. The coefficient for record linkage is now statistically significant, implying that our results are more conservative and that we can be more confident about them.

A fourth robustness check addresses the concern that interviewers' reading time of introductory items is not shortened per se but rather shifted to the duration of the subsequent item as interviewers remember the wording of the introduction text and click "continue" before finishing the verbal introduction. Then our measure would reflect only one part of the reading time and imply an increase in the duration of the subsequent item over the field period. However, our results do not support this concern. The durations of subsequent items either remain stable or decrease over the field period and therefore support our main results of decreasing reading times (supplement 2.4 of the online [Supplementary Material](#)).

A further concern is directed toward interviewers' prior experience. Learning might not only take place over one field period but begin much earlier. In this respect, overall interviewer experience has been shown to be negatively associated with interview length over the field period ([Olson and Peytchev 2007](#)) but not with question response times for each question overall ([Olson and Smyth 2015](#)). To address the potential moderation effect of interviewer experience, we included overall interviewer experience in years and prior experience with SHARE as interaction terms (see supplements 2.5 and 2.6 of the online [Supplementary Material](#)). Overall, the moderating effects are mostly significant but rather small compared with the total decrease in reading time. In all cases, they tend to attenuate the negative effect (i.e., the learning effect is less strong for experienced interviewers) mainly because of their lower starting level compared with inexperienced interviewers.

3.3 Does It Matter? Consequences of Decreasing Reading Time on Respondents' Response Behavior

In the second stage of our study, we analyzed the extent to which the observed decline in reading time substantially affects respondents' response behavior. Therefore, we used the reading times of the selected introductory items as

Table 2. Effect of Changes in Interviewers' Reading Time on Respondents' Survey Responses

Introductory item	Survey response item	Average change in reading time in seconds	Effect of average change in % points (SE)
Intro to overall SHARE interview	Refusal to income	−5.7	−0.16 (0.19)
Intro to record linkage	Consent to record linkage	−58.6	−7.01*** (1.23)
Intro to health care expenditures	Paid anything out-of-pocket	−11.8	−1.08** (0.35)
Intro to local area information	Residential satisfaction	−8.2	−0.06 (0.11)
Intro to recall test	Amount of words recalled	−18.8	−2.34** (0.82)
Intro to chair stand	Compliance with chair stand test	−17.4	−2.86*** (0.30)

NOTE.—Each line represents one fixed effects model on the respective survey response with reading time of intro text as explanatory variable and days in field, respondent characteristics, and interview specifics as controls. Average change in percentage points is based on average marginal effects of logistic fixed effects regressions for refusal to income, consent to record linkage, paid anything out-of-pocket, and compliance with chair stand test. Linear fixed effects models were applied for residential satisfaction and amount of words recalled.

Weighted results. Panel-robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

main explanatory variables, while our dependent variables were the survey answers described previously. We again ran fixed effects models and controlled for period effects, respondent characteristics, and the interview situation. The results are presented in [table 2](#). See online supplement 3 for full regression table including all control variables.

It can be seen that the average decrease in reading out the record linkage introduction between the first and the last interview of each interviewer is approximately 59 seconds in our sample. This huge decrease is most likely due to not giving the respondents (in particular the partner if there are two interviews in the household) sufficient time to carefully read the SHARE statement about privacy concerns. If we wish to know what this decrease means for a connected survey response such as the consent rate, we must multiply this value with the regression coefficient (or the average marginal effect in a logistic model) obtained in the fixed effects model. The result of this calculation gives a coefficient of -0.07 . Thus, the average decrease in reading time with respect to the record linkage introduction significantly reduces the consent rate, on average,

Table 3. Predicted Values 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	0.07	0.07
Consent to record linkage	0.64	0.79***
Paid anything out-of-pocket	0.42	0.50***
Residential satisfaction	0.18	0.19*
Amount of words recalled	5.33	5.40*
Compliance with chair stand test	0.89	0.96***

NOTE.—Predicted values are based on marginal effects after running item-specific logistic fixed effects models on refusal to income, consent to record linkage, paid anything out-of-pocket, and compliance with chair stand test and linear fixed effects models on residential satisfaction and amount of words recalled. Weighted results.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

by seven percentage points and thus supports our assumption (see the last column of [table 2](#)).

Similar findings are reported with respect to the decrease in reading the introductions for the cognitive recall test and the chair stand measurement. Thus, the average decrease of 17.4 seconds in the latter case results in a reduced compliance rate of 2.9 percentage points. Although these effects explicitly concern within-survey requests, there is also a significant effect with respect to the definition of out-of-pocket payments. Here, a shorter reading time results in stating fewer payments. However, this effect is rather small (minus one percentage point). The other survey responses (refusal to income question and a less reliable indication of residential satisfaction) are not changed significantly by a decreased duration of the introductory item. This supports our assumption of different effect sizes being dependent on the amount of informational content that has (not) been read out by the interviewer.

Thus far, we have restricted our analyses to the effects of (within) changes in reading time. We complemented this within-perspective by looking at the absolute level of interviewers’ reading times and how differences here might affect the overall level of respondents’ response behavior. This complement analysis provides a better intuition for the magnitude of the observed deviations from standardized interviewing. [Table 3](#) reports the predicted values for two different levels of reading time, illustrating the effect of those interviewers who simply click through an introductory text (10 percent level) compared with those who read out the introductory texts in a standardized way (90 percent level).

Substantial differences can again be observed for consent to record linkage (fifteen percentage points), out-of-pocket payment (nine percentage points), and compliance to the chair stand test. If, for example, all interviewers took only one second (10 percent level) to read the introductory item to the chair stand test, 89.4 percent of the respondents would comply. If, instead, all interviewers took 72 seconds (90 percent level), the compliance rate would be 95.9 percent, a difference of 6.5 percentage points. The effect for the other outcomes is not substantial, although the difference for variance in responses on residential satisfaction (one percent point) and the number of words recalled (.07 words or 1.3 percentage points) are statistically significant. Overall, our findings suggest that reading time matters in particular for within-survey requests, but also—although to a lesser degree—for most other survey responses under investigation.

4. DISCUSSION

To obtain accurate survey estimates, the interviewer is of particular interest within the total survey error framework. The common approach to ensure comparability and to minimize interviewer influence on measurement error is standardized interviewing, which includes reading questions and introductions as worded as one key component. Our main finding is that interviewers' reading time of introductory items significantly decreases over the survey's field period, which contrasts with the goal of standardized interviewing. After carefully controlling for potential time-variant confounders, we are confident in ascribing a negative effect of the increasing order of conducted interviews to an interviewer's learning behavior to shorten, skip, or speed through the item. This pattern is comparable across several items and nearly all countries regardless of language differences and thus can be generalized. Hence, our first two hypotheses can be confirmed.

In this article, we further answered the crucial question of whether this decrease in reading time has substantial consequences for the response behavior of respondents. In this regard, our analyses reveal a rather large effect of decreasing reading time on within-survey requests, such as the consent given by respondents to link their survey answers to administrative data or take cognitive and physical tests. Our findings suggest that respondents are sensitive to confidentiality concerns and their physical abilities. When confronted with a hastily or partially introduced request, they frequently refuse to consent. In addition to these within-survey requests, other variables show much smaller effects. For survey researchers, this outcome is good news. However, our results should not lead to the conclusion that strong deviations from standardized interviewing can be largely neglected. Rather, it is more likely that the size of effects strongly depends on the amount of information which is present in the respective introductory text, as assumed in hypothesis 3, which can also be confirmed.

In addition, our findings have implications for various areas. First, one major implication concerns questionnaire design. It is well-known that the visual presentation of questions and answers can influence the way information is read and whether it is read at all (e.g., [Dillman, Smyth, and Christian 2014](#)). Usually, however, questionnaire design only takes respondents into account, neglecting the importance of interviewers in this context. For example, introductions are usually presented in block texts. However, such a format requires more concentration in reading them out completely and increases the probability of skipping. A clearer structuring of introductory texts might be a useful improvement.

Another important implication is related to fieldwork management and monitoring. Overall, the interviews utilized for this study are relatively long, and interviewers are paid by completed interview. Although we could not test an effect of payment schemes due to lack of variation, shortening interviews seems to be a rational behavior for interviewers. As payment-per-interview schemes are usually applied in large-scale surveys (at least in Europe) and survey organizations have a strong incentive to maintain this routine in terms of minimizing their management costs, changes will be difficult. Thus, although randomized payment schemes for interviewers would be optimal in terms of analyzing interviewer's reading behavior, this is not very realistic. One possible starting point, however, might be to add steeper bonus and/or malus systems that reward good interviewers and punish delinquent ones. In this respect, it is important to emphasize that good interviewer behavior should be measured not only in terms of response rates but based on the interviewer's actual behavior in interacting with respondents. We argue that monitoring interviewers more closely is necessary to keep their work quality high. We know from the Italian case that interviewers perform well when receiving feedback on their interview length and reading time.

Closely connected with this finding, a third implication for interviewer training is observed. SHARE uses a centralized train-the-trainer system to ensure standardized and harmonized training across all countries ([Börsch-Supan and Jürges 2005](#)). Prior to fieldwork, special emphasis should be placed on training the introduction of within-survey requests. As training impact tends to diminish over time, retraining during the survey's field period might help reinforce good interviewing practices, especially for interviewers who conduct many interviews. This is important, as our analyses revealed that these interviewers show very short reading times on average.

As with any study, our analysis has limitations. First, although we carefully selected items that should measure sole interviewer behavior and further controlled for comprehension problems by respondents, we cannot completely rule out verbal interactions between respondent and interviewer as we have no behavior-coded data. However, by using introductory items rather than classical survey questions, we actually try to minimize this problem as much as possible. Hence, the measures of an interviewer's reading behavior used are less

skewed than indicators that do not allow differentiating between the respective share of interviewers and respondents. Our approach is thus more precise than using the overall interview length, as applied in previous studies.

Second, we cannot disentangle mechanisms leading to shorter reading times. Our data does not reveal if a change in the duration is due to shortening, skipping, changing words in the introductory item, or simply reading it more fluently. Thus, experimental designs, preferably in combination with the (audio) recording of a selection of interviews, are strongly encouraged (see [Ackermann-Piek and Massing 2014](#); [Zabal, Martin, Massing, Ackermann, Helmschrott, et al. 2014](#)). This approach would be particularly useful for within-survey requests where heterogeneity in the interviewer-respondent interaction might be possible. However, audio recordings demand careful consideration of country-specific legal regulations and requirements for consent. The degree to which this might lead to further nonresponse is unclear, and potentially negative consequences would need to be evaluated before applying such measures. In this respect, our study presents a promising approach to analyzing deviant interviewer behavior in the common situation when (audio) recordings are not available.

Third, although we 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 (time-variant) variables. However, our analytical strategy of applying fixed effects models instead of the more commonly used multilevel models has the advantage that all time-constant influences are already implicitly controlled for and cannot bias the estimates. Together with our main interest in changes of interviewers' reading behavior over the survey's field period, this model decision is a sensible contribution to previous research in this area, which has focused more on differences between interviewers from a cross-sectional perspective.

Fourth, in terms of the data used, one must take into account that the target population in SHARE comprises people age fifty and older rather than the general population. As mentioned before, it is possible that older people are more strongly affected by decreasing reading times, (e.g., if they already have problems understanding certain questions or following the interview procedure in general). In this respect, we consider it important to extend our analyses to other parts of the population by using additional data in future research. Similarly, studies possessing reliable information on interviewer-respondent continuity should consider changes over several waves, as changes in interviewer training or monitoring strategies might also influence interviewers' reading behavior.

Supplementary Materials

[Supplementary materials](#) are available online at academic.oup.com/jssam.

Appendix

Table A. 1. Wording of Introductory Texts

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. <i>1. Continue</i>
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. <i>1. Continue</i>
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. <i>1. Continue</i>
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. <i>1. Continue</i>
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 <i>1. Continue</i>

Continued

Table A.1. *Continued*

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 Non-proxy section. No proxy allowed. If the respondent is not capable of answering any of these questions on her/his own, press CTRL-K at each question. <i>1. Continue</i>

NOTE.—SOURCE: SHARE Wave 5 questionnaire.

Table A.2. Wording of Survey Response Variables

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 to re- cord linkage	LI003_Consent <i>IWER: Did R consent to the record linkage?</i> <i>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 addressed envelope and bring it to the mailbox.</i> 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.
Paid 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

Continued

Table A.2. Continued

Vandalism or crime	HH023_LocalVandalism Vandalism or crime is a big problem in this area. (Would you say you strongly agree, agree, disagree or 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?)												
Amount of words recalled	CF104_Learn1 Now please tell me all the words you can recall. <table><tr><td>1. Hotel</td><td>5. Gold</td><td>9. King</td></tr><tr><td>2. River</td><td>6. Market</td><td>10. Book</td></tr><tr><td>3. Tree</td><td>7. Paper</td><td>96. None of these</td></tr><tr><td>4. Skin</td><td>8. Child</td><td></td></tr></table> Interviewer enters the words the respondent correctly recalls. 4 sets of wording lists are assigned randomly (CF104_Learn1-CF107_Learn1).	1. Hotel	5. Gold	9. King	2. River	6. Market	10. Book	3. Tree	7. Paper	96. None of these	4. Skin	8. Child	
1. Hotel	5. Gold	9. King											
2. River	6. Market	10. Book											
3. Tree	7. Paper	96. None of these											
4. Skin	8. Child												
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? <table><tr><td>1. Yes</td></tr><tr><td>5. No</td></tr></table>	1. Yes	5. No										
1. Yes													
5. No													

NOTE.—SOURCE: SHARE Wave 5 questionnaire.

Table A.3. Descriptive Statistics for Interviews and Interviewers' Reading Behavior

		Mean reading time of introductory texts in seconds						
N (obs.)	N (int.)	Mean number of interviews per interviewer	Overall SHARE interview	Record linkage	Health care expenditure	Local area information	Recall test	Chair stand
Austria	4,252	87	11.62 (16.24)		17.91 (16.62)	9.32 (9.19)	18.59 (20.18)	17.57 (22.89)
Belgium	5,473	145	11.44 (13.79)		16.15 (15.10)	14.60 (11.05)	28.20 (22.72)	33.35 (38.56)
Czech Republic	5,468	153	13.01 (18.52)		9.98 (12.95)	11.83 (9.87)	20.18 (21.96)	15.33 (20.41)
Denmark	4,068	57	15.53 (20.21)		11.33 (9.95)	14.62 (9.65)	31.77 (19.87)	29.40 (37.60)
Estonia	5,576	96	9.85 (12.89)		7.73 (9.91)	9.18 (9.41)	18.79 (19.43)	18.46 (25.34)
France	4,426	176	10.22 (11.82)		15.96 (14.33)	13.77 (9.96)	34.33 (26.40)	40.75 (40.24)
Germany	5,662	177	14.49 (15.92)	79.55 (75.86)	16.43 (14.00)	17.16 (11.64)	42.64 (29.50)	24.67 (25.32)
Israel	1,901	21	8.54 (15.12)		3.44 (5.18)	2.96 (3.30)	11.93 (14.08)	2.28 (4.45)
Italy	4,588	142	10.96 (15.88)		17.18 (22.83)	8.96 (9.58)	20.41 (23.42)	13.07 (19.12)
Netherlands	4,052	127	12.62 (17.26)		13.43 (13.10)	11.38 (8.98)	32.27 (26.01)	35.22 (39.88)
Luxembourg	1,556	61	15.03 (14.83)		17.66 (16.53)	15.46 (10.57)	34.33 (24.70)	25.53 (28.27)
Slovenia	2,687	45	5.15 (8.21)		3.50 (4.80)	5.41 (6.25)	14.80 (14.70)	12.21 (19.15)
Spain	6,213	140	9.18 (14.49)		11.75 (17.92)	6.68 (8.72)	17.46 (21.87)	10.85 (19.04)
Sweden	4,469	94	17.76 (24.57)		12.20 (10.11)	17.14 (11.35)	31.21 (21.77)	26.33 (34.08)
Switzerland	3,006	60	10.95 (13.61)		13.80 (13.75)	13.17 (10.14)	25.71 (21.34)	29.29 (39.73)
Total	63,397	1,581	12.17 (16.73)	79.55 (75.86)	12.93 (14.85)	11.80 (10.50)	25.85 (24.16)	22.70 (31.06)

NOTE.—N (obs.) = Number of observations/interviews, N (int.) = Number of interviewers. Please note that the number of observations per introductory text differs due to routing. Standard deviations in parentheses.

Table A.4. Descriptive Statistics for Key Survey Response Variables

Variable	Mean	SD	Min	Max	Description
Refusal to income	0.07	0.25	0	1	Refuse to provide information on income
Consent to record linkage	0.73	0.45	0	1	Consent to record linkage
Paid anything out-of-pocket	0.45	0.50	0	1	Paid anything out-of-pocket for doctoral visits
Residential satisfaction	0.18	0.13	0	0.71	Variance of four items measuring residential satisfaction with the local area (feeling part, vandalism or crime, cleanliness, help available)
Amount of words recalled	5.32	1.84	0	10	Number of words recalled in 10 words learning test (first recall)
Compliance with chair stand test	0.84	0.37	0	1	Compliance with chair stand measurement

NOTE.—DATA: SHARE Wave 5. Unweighted data. SD = Standard Deviation.

Table A.5. Descriptive Statistics for Control Variables

Variable	Mean	SD	Description
<i>Period effects</i>			
Days in field	122.95	77.86	The number of days the country is in the field conducting interviews
<i>Respondent characteristics</i>			
Female	0.56	0.50	Gender of respondent (0=male, 1=female)
Age	66.64	10.22	Age of respondent (range: 22 to 104)
Years of education	11.18	4.33	Years of education (range: 0 to 25)
Employment status			Current employment status
retired	0.56	0.50	
(self-) employed	0.28	0.45	Employed or self-employed
other	0.16	0.36	
Household income			Total income received by all household members in the last month
1 st Quartile	0.20	0.40	
2 nd Quartile	0.20	0.40	
3 rd Quartile	0.20	0.40	
4 th Quartile	0.20	0.40	
Missing	0.19	0.39	
Household size	2.14	0.95	Number of persons living in the household (range: 1 to 11)

Continued

Table A.5. *Continued*

Variable	Mean	SD	Description
Urban area	0.42	0.49	Living in a large town, suburbs or outskirts of a big city or a big city (reference is rural=0, i.e., small town, rural area, or village)
Self-perceived health	0.47	0.27	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	0.17	0.37	Having one or more limitations with instrumental activities of daily living (iadl)
Grip strength	32.27	13.57	Maximum grip strength measurement of respondent (range: 0 to 90)
Social activities	0.87	0.34	Participation in any social activity during the last year (voluntary work, sports, politically active, educational course, playing cards, games)
Good/very good willingness to respond	0.90	0.30	Willingness to respond to interview was coded by interviewer (reference: fair/bad willingness to respond)
Never/almost never asked for clarification	0.76	0.43	Clarification on questions was coded by interviewer (reference: now and then, (often, very often, always)
Very often/always understood questions	0.76	0.43	Understanding of questions was coded by interviewer (reference: never, almost never, now and then, often)
No help needed with showcards	0.92	0.28	Help with reading showcards was coded by interviewer (reference: help needed)
Times participated	2.28	1.37	Number of times the respondent participated in SHARE (range: 1 to 5)
Interrupted pattern	0.12	0.33	Interrupted participation pattern in SHARE survey (reference is constant participation)
<i>Interview specifics</i>			
Longitudinal interview	0.65	0.48	Interview version (reference: baseline interview)
Late in sample	0.10	0.30	Within the last 10% of conducted interviews in sample
Second respondent	0.32	0.46	Interview of second respondent in household (yes/no)
Interviews per day	1.82	1.16	Number of interviews started per interviewer within one day (range: 1 to 13)

NOTE.—DATA: SHARE Wave 5. Unweighted data. SD=Standard Deviation.

References

- AAPOR (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*, AAPOR.
- Ackermann-Piek, D., and N. Massing (2014), "Interviewer Behavior and Interviewer Characteristics in PIAAC Germany," *Methods, Data, Analyses*, 8, 199–222.
- Bergmann, M., T. Kneip, G. De Luca, and A. Scherpenzeel (2017), "Survey Participation in the Survey of Health, Ageing and Retirement in Europe (Share), Wave 1-6," SHARE - Working Paper Series 31-2017, Munich Center for the Economics of Aging (MEA).
- Beullens, K., and G. Loosveldt (2016), "Interviewer Effects in the European Social Survey," *Survey Research Methods*, 10, 103–118.
- Böhme, M., and T. Stöhr (2014), "Household Interview Duration Analysis in CAPI Survey Management," *Field Methods*, 26, 390–405.
- Börsch-Supan, A. (2015), Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release Version: 6.0.0. Data set. SHARE-ERIC. DOI: 10.6103/SHARE.w5.600.
- Börsch-Supan, A., M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber (2013), "Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (Share)," *International Journal of Epidemiology*, 42, 992–1001.
- Börsch-Supan, A., and H. Jürges (2005), *The Survey of Health, Ageing and Retirement in Europe – Methodology*, Mannheim: MEA.
- Bristle, J. (2015), "Measuring Interview Length with Keystroke Data," in *SHARE Wave 5: Innovations & Methodology*, eds. F. Malter, and A. Börsch-Supan, 165–176, Munich: MEA, Max Planck Institute for Social Law and Social Policy.
- Brüderl, J., B. Huyer-May, and C. Schmiedeberg (2012), "Interviewer Behavior and the Quality of Social Network Data," in *Survey Standardization and Interviewers' Deviations—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, 114, Frankfurt: Peter Lang.
- Cannell, C.F., K.H. Marquis, and A. Laurent (1977), "A Summary of Studies of Interviewing Methodology", *Vital and Health Statistics*, 2, 1–78.
- Chromy, J. R., J. Eyerman, D. Odom, M. E. McNeeley, and A. 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*, eds. J. Kennet, and J. Gfroerer, 59–87, Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.
- Clark, T. S., and D. A. Linzer (2015), "Should I Use Fixed or Random Effects?" *Political Science Research and Methods*, 3, 399–408.
- 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, 271–286.
- Davis, R. E., M. P. Couper, N. K. Janz, C. H. Caldwell, and K. Resnicow (2010), "Interviewer Effects in Public Health Surveys," *Health Education Research*, 25, 14–26.
- De Luca, G., C. Rossetti, and F. Malter (2015), "Sample Design and Weighting Strategies in Share Wave 5," in *SHARE Wave 5: Innovations & Methodology*, eds. F. Malter, and A. Börsch-Supan, 75–84, Munich: MEA, Max Planck Institute for Social Law and Social Policy.
- Dillman, D. A., J. D. Smyth, and L. M. Christian (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, Hoboken, NJ: John Wiley & Sons.
- ESS7 (2016), *ESS7 - 2014 Documentation Report*, London: ESS ERIC.
- Fowler, F. J. (1991), "Reducing Interviewer-Related Error Through Interviewer Training, Supervision, and Other Means," in *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, 259–278, Hoboken, NJ: John Wiley & Sons.
- Fowler, F. J., and T. W. Mangione (1990), *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*, Newbury Park, CA: SAGE Publications.
- Groves, R. M., and L. Lyberg (2010), "Total Survey Error. Past, Present, and Future," *Public Opinion Quarterly*, 74, 849–879.

- Groves, R. M., and L. J. Magilavy (1986), "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," *Public Opinion Quarterly*, 50, 251–266.
- Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau (2009), *Survey Methodology*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Haan, M., Y. Ongena, and M. Huiskes (2013), "Interviewers' Question: Rewording Not Always a Bad Thing," in *Interviewers' Deviations in Surveys: Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, 173–193, Frankfurt a. M.: Peter Lang.
- Holbrook, A. L., T. P. Johnson, Y. I. Cho, S. Shavitt, N. Chavez, and S. Weiner (2016), "Do Interviewer Errors Help Explain the Impact of Question Characteristics on Respondent Difficulties?," *Survey Practice*, 9(2), <http://www.surveypactice.org/article/2818>
- Hyman, H. H., and Cobb, W., J. (1954). *Interviewing in Social Research*. Chicago, IL: University of Chicago Press.
- Josten, M., and M. Trappmann (2016), "Interviewer Effects on a Network-Size Filter Question," *Journal of Official Statistics*, 32, 349–373.
- Kirchner, A., and Olson, K. (2017), "Examining Changes of Interview Length over the Course of the Field Period," *Journal of Survey Statistics and Methodology*, 5, 84–108.
- Kish, L. (1962), "Studies of Interviewer Variance for Attitudinal Variables," *Journal of the American Statistical Association*, 57, 92–115.
- Korbmacher, J. M., and M. Schroeder (2013), "Consent When Linking Survey Data with Administrative Records: The Role of the Interviewer," *Survey Research Methods*, 7, 115–131.
- Kosyakova, Y., J. Skopek, and S. Eckman (2015), "Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach," *International Journal of Public Opinion Research*, 27, 417–431.
- Kreuter, F. (2013), *Improving Surveys with Paradata: Analytic Uses of Process Information*, Hoboken, NJ: John Wiley & Sons.
- 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, 69–78.
- Lyberg, L. (2012), "Survey Quality," *Survey Methodology*, 38, 107–130.
- Malter, F. (2013), "Fieldwork Monitoring in the Survey of Health, Ageing and Retirement in Europe (SHARE)," *Survey Methods: Insights from the Field*. doi:10.13094/SMIF-2014-00006.
- Mangione, T. W., F. J. Fowler, and T. A. Louis (1992), "Question Characteristics and Interviewer Effects," *Journal of Official Statistics*, 8, 293–307.
- Matschinger, H., S. Bernert, and M. C. Angermeyer (2005), "An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview," *Journal of Official Statistics*, 21, 657–674.
- Olson, K., and I. Bilgen (2011), "The Role of Interviewer Experience on Acquiescence," *Public Opinion Quarterly*, 75, 99–114.
- Olson, K., and A. Peytchev (2007), "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes," *Public Opinion Quarterly*, 71, 273–286.
- Olson, K., and J. D. Smyth (2015), "The Effect of CATI Questions, Respondents, and Interviewers on Response Time," *Journal of Survey Statistics and Methodology*, 3, 361–396.
- Ongena, Y. P. (2003), "Pre-Testing the ESS-Questionnaire Using Interaction Analysis," in *European Social Survey CCT Meeting*, Sociaal en Cultureel Planbureau, Den Haag.
- Ongena, Y. P., and W. Dijkstra (2007), "A Model of Cognitive Processes and Conversational Principles in Survey Interview Interaction," *Applied Cognitive Psychology*, 21, 145–163.
- Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.
- Sakshaug, J. W. (2013), "Using Paradata to Study Response to within-Survey Requests," in *Improving Surveys with Paradata*, ed. F. Kreuter, 171–190, Hoboken, NJ: John Wiley & Sons.
- Schaeffer, N. C., J. Dykema, and D. W. Maynard (2010), "Interviewers and Interviewing," in *Handbook of Survey Research*, eds. P. V. Marsden, and J. D. Wright, 437–470, Howard House, UK: Emerald Group Publishing Limited.
- Schnell, R., and F. Kreuter (2005), "Separating Interviewer and Sampling-Point Effects," *Journal of Official Statistics*, 21, 389–410.
- Schober, M. F., and F. G. Conrad (1997), "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly*, 61, 576–602.

- West, B. T., and A. G. Blom (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5, 175–211.
- West, B. T., F. G. Conrad, F. Kreuter, and F. Mittereder (2018), "Nonresponse and Measurement Error Variance Among Interviewers in Standardized and Conversational Interviewing," *Journal of Survey Statistics and Methodology*, 6, 335–359.
- West, B. T., F. Kreuter, and U. Jaenichen (2013), "Interviewer" Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?" *Journal of Official Statistics*, 29, 277–297.
- Wooldridge, J. (2013), *Introductory Econometrics: A Modern Approach* (5th ed.), Mason, OH: South-Western Cengage Learning.
- Yan, T., and K. Olson (2013), "Analyzing Paradata to Investigate Measurement Error," in *Improving Surveys with Paradata. Analytic Uses of Process Information*, ed. F. Kreuter, 73–95, Hoboken, NJ: John Wiley & Sons.
- Zabal, A., S. Martin, N. Massing, D. Ackermann, S. Helmschrott, I. Barkow, and B. Rammstedt (2014), *PIAAC Germany 2012: Technical Report*, Münster: Waxmann.