Longitudinal and Life Course Studies • vol 10 • no 4 • 513–530
 • Paper received November 2018/Revised June 2019/Published online October 2019
Online ISSN 1757-9597 • https://doi.org/10.1332/175795919X15694136530293

Special Issue: Methodological Issues in Longitudinal Surveys

RESEARCH ARTICLE

Preventing interview falsifications during fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE)

Michael Bergmann[1], *bergmann@mea.mpisoc.mpg.de*
Karin Schuller, k.schuller@mea.mpisoc.mpg.de
Frederic Malter, malter@mea.mpisoc.mpg.de
Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy, Germany

The fabrication of an entire interview, is a rare event in the Survey of Health, Ageing and Retirement in Europe (SHARE) but can nevertheless lead to negative consequences regarding the panel sample, such as a loss in sample size or the need for time-consuming data corrections of information collected in previous waves. The work presented in this article started with the discovery of a case of interviewer fabrication after fieldwork for the sixth wave of SHARE was completed. As a consequence, we developed a technical procedure to identify interview fabrication and deal with it during ongoing fieldwork in the seventh wave. Unlike previous work that often used small experimental datasets and/or only a few variables to identify fake interviews, we implemented a more complex approach with a multivariate cluster analysis using many indicators from the available CAPI data and paradata. Analyses with the known outcome (interview fabrication or not) in wave 6 revealed that we were able to correctly identify a large number of the truly faked interviews while keeping the rate of 'false alarms' rather low. With these promising results, we started using the same script during the fieldwork for wave 7. We provided the survey agencies with information for targeted (instead of random) back checks to increase the likelihood of confirming our initial suspicion. The results show that only a very small number of interview fabrications could be unequivocally identified.

Key words interview falsification • interviewer behaviour • cluster analysis • panel • paradata

To cite this article: Bergmann, M., Schuller, K. and Malter, F. (2019) Preventing interview falsifications during fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE), *Longitudinal and Life Course Studies*, vol 10, no 4, 513–530, DOI: 10.1332/175795919X15694136530293

Introduction

Interviewer falsification ('fake interviews') is a problem in all interviewer-conducted surveys. While there are many variations and different reasons for interviewers deviating from properly administering the survey (for an overview see Murphy et al, 2016), here we will only deal with the most extreme form of deviation: interviewers' fabrication of entire interviews. For the sake of brevity, we will refer to this issue with the term 'curbstoning', coined by the U.S. Census Bureau as 'curbstoning' (Werker, 1981; Ericksen and Kadane, 1985). In the original sense, curbstoning refers to 'sitting on a curbstone and completing questionnaires, rather than interviewing respondents' (Koczela et al, 2015: 414). In contrast, we will not address any ideas on minimum interview quality, for example minimum levels or constellations of data that would constitute a cut-off for releasing an interview or not. The latter is very difficult to conceptualise and far less enforceable in complex (panel) surveys.

The consequences of curbstoning for substantive results were found to be non-negligible in the case of multivariate analyses (such as Schräpler and Wagner, 2003). Hence, simply ignoring the issue of curbstoning may harm the scientific credibility of any survey (Werker, 1981, for example). Although curbstoning is believed to be a rare event (Li et al, 2011; Murphy et al, 2016; Schupp, 2018), there is evidence that it occurs not only in commercial surveys with limited ranges but also in large-scale studies with strong scientific guidelines (see Hood and Bushery, 1997; Schräpler and Wagner, 2003; Blasius and Thiessen, 2015). The share of (identified) faked interviews is usually below 5% in cross-sectional studies (Koch, 1995; Hood and Bushery, 1997), but it has been found to be drastically higher in some cases (Turner et al, 2002; Murphy et al, 2004). Curbstoning is less frequent in panel studies, with a share of 0.1%–2.4% in the Socio-Economic Panel (Schräpler, 2004) or 0.4% in the Current Population Study and the National Crime Survey (Schreiner et al, 1988).

In our view, it still seems important to take a closer look at curbstoning – in particular concerning panel surveys, as complications and negative consequences can accumulate. First, in panel surveys, information on a household from previous waves can be preloaded. Thus, conducting a proper panel interview with 'never-interviewed households' is nearly impossible when this preloaded information cannot be verified with the present respondent(s). Second, such a method of revealing fabricated interviews involves time-consuming data corrections of information collected during previous waves. Frequently, the only way to guarantee high data quality is to retrospectively delete complete interviews, with the effect that the study will successively fail to achieve its primary aim: analysing individual change over time. Therefore, fabricated interviews should ideally be detected as early as possible, preferably during ongoing fieldwork to enable timely reinterviews of the correct respondents. This would help reduce cumbersome corrections of fabricated interviews and, at the same time, increase data quality in the long run. While curbstoning seems to be more problematic in panel surveys regarding its consequences for data management and substantial analyses, the opportunities to identify fabricated interviews are also greater in panel surveys than in cross-sectional studies because panel data offer the

possibility to compare interview results with those of previous waves to identify inconsistencies.

Methods of detecting curbstoning involve the comparison of demographic information of the interviewed persons with data from registration offices (Koch, 1995), the analysis of response behaviour within the data (Murphy et al, 2004; Schäfer et al, 2005; Yamamoto and Lennon, 2018) or the analyses of metadata (Hood and Bushery, 1997; Turner et al, 2002; Murphy et al, 2004; Yamamoto and Lennon, 2018). Another method for detecting curbstoning is the examination of the distribution of the first digit of all numbers in metric answers, to test for compliance with Benford's law (Schräpler, 2010). The most common method of identifying curbstoning, however, is by recontacting interviewed households and asking them if they have actually participated in the interview (Murphy et al, 2016). It is not possible to reinterview the entire sample for financial and administrative reasons, so a selection from the sample is usually chosen. This selection can be performed randomly or based on several indicators to create a focused reinterview pool (Hood and Bushery, 1997; Turner et al, 2002; Menold et al, 2013). In this respect, the combination of recontacting a focused sample and conducting random back checks has been found to be the most effective way of identifying curbstoning (Bredl et al, 2013). The literature notes different strategies for selecting a focused sample of interviews that should be recontacted (for an overview see Bredl et al, 2013). In the scholarly debate on curbstoning, it is assumed that fabricated data from falsifying interviewers deviate from data obtained from real respondents in certain aspects. Possible indicators can be gained from computer-assisted personal interviewing (CAPI) data (such as follow-up questions, item non-response) or from paradata that is frequently available as a by-product of the data collection process (interview length or performance indicators, for example).

In this respect, our primary goal was to develop strategies for SHARE to generate such a focused sample and to equip survey agencies with a more informed sample of suspicious interviews from interviewers that need to be checked. To achieve this, we implemented a technical procedure to detect curbstoning during the ongoing fieldwork in wave 7. We built on the results of previous research by using several indicators that have been shown to be important in detecting fake interviews (see, for example, Menold et al, 2013; Murphy et al, 2016). These studies frequently had the shortcomings that either the method could not be tested on real survey data but only on experimental data that was intentionally falsified for that purpose (Menold et al, 2013) or that only very few indicators could be used (Bredl et al, 2012). In this context, we contribute to the literature by conducting a multivariate cluster analysis with a large set of indicators from CAPI data and paradata and panel information about respondents' answers from previous waves to identify curbstoning in a real survey setting. In this article, we start with a brief overview of previous findings from which we derive testable hypotheses (a more exhaustive review of indicators can be found in Murphy et al, 2016). Afterwards, we describe the implementation of the cluster analysis during fieldwork in waves 6 and 7 and how it worked out in each case. We conclude with a summary of our findings and a discussion of lessons learned for future research in this area.

Theoretical considerations and hypotheses regarding interview falsifications

Theoretical assumptions regarding the choice of indicators to build a focused sample to be back checked – if applied at all – are frequently based on the satisficing model developed by Krosnick and Alwin (1987), postulating that respondents will usually minimise their (cognitive) effort and hence choose the first acceptable answer when responding to survey questions. The basic idea of satisficing has further evolved in the context of interview falsifications, assuming that falsifying interviewers also want to save time and effort, while at the same time, they try to minimise the risk of being detected (Japec, 2006; Menold et al, 2013). From current research we know, for instance, that falsifiers use the 'other' option in the questionnaire less often than respondents would in a real setting to avoid the cognitive burden of answering semi-open-ended questions (Menold et al, 2013). The same logic applies to filter questions that frequently have the format of a 'yes or no' query and are meant to help respondents avoid answering more detailed follow-up questions that do not pertain to them (see Allen, 2017). In this respect, Hood and Bushery (1997) show that falsifiers avoid follow-up questions to save time and effort, and Shaeffer et al (2005) present evidence that falsifiers show less variation in choosing answer options in multi-item scales; hence, more so-called 'straight-lining' becomes apparent. Regarding this argumentation, interviewers should also have a shorter interview length. However, previous research does not come to a clear conclusion on this aspect. While Schreiner et al (1988) showed that interviewers with shorter interviews are more likely to fabricate data, Bushery et al (1999) and Murphy et al (2004) found that falsifiers report a remarkably long *or* short time to complete the entire questionnaire or certain modules. Additionally, interviewers who complete many questionnaires within a given time period are supposed to be at a higher risk for curbstoning (Murphy et al, 2016). In addition, Hood and Bushery (1997) observed that falsifiers have an elevated rate of households that have been labelled ineligible or without telephone numbers. Moreover, there is evidence that fieldwork agencies operating a survey might falsify their data by simply producing (near) duplicates of interviews to save time and effort (Blasius and Thiessen, 2013, 2015; Kuriakose and Robbins, 2016).

While all these findings are in accordance with the satisficing hypothesis originally referring to respondent behaviour, interviewers who fabricate their interviews try to avoid detection in addition, resulting in less satisficing with respect to certain contexts (Menold et al, 2013). The reason for this is that data falsifiers do not want to become conspicuous by producing too many non-substantive answers; however, at the same time, they are not able to produce the same amount of variability as that found in real data when using stereotypes of 'typical' respondents. Hence, data falsifiers frequently show a lower level of item non-response, less often choose the first answer option (less primacy effects) and have overall less extreme answer patterns than honest interviewers (Shaeffer et al, 2005; Bredl et al, 2012; Menold et al, 2013). The same rationale of avoiding detection should hold for interviewers pretending to have used proxy respondents (someone who answers the questions in the questionnaire instead of the selected respondent) or to have done rather lengthy physical tests, such as measurements of hand grip strength by simply putting in preferably realistic numbers. In these cases, less satisficing (for example, using numbers that are not rounded to multiples of five or ten) does not require additional effort but rather gives the impression of inconspicuous data.

Based on this brief summary of the previous findings, we expect that interviewers fabricating their interviews want to save time and effort. Therefore, we assume that they have a lower number of contact attempts, fewer interviewer notes, a shorter interview duration, a lower number of asked items, fewer 'other' or 'code all that apply' answers and fewer follow-up questions. Similarly, we expect more duplicate interviews and more straight-lining (providing the same answer to all questions for a block of questions with identical answer categories; see Kaminska et al, 2010) because this also saves time and effort. On the other hand, we expect that falsifiers have a lower item non-response rate, less extreme answer patterns and a lower level of primacy effects (an increased likelihood of selecting an answer category when it is placed at the beginning of a list; see Krosnick and Alwin, 1987) compared to honest interviewers because they want to avoid detection. In line with this, we assume that data falsifiers generate more proxy interviews and more grip strength measurements. In this respect, we also expect less rounding of numeric values because putting in an invented but realistic number for the grip strength test does not take any additional effort. As a by-product of this behaviour but also because of the dominant payment structure of European survey agencies that disburse payments to their interviewers per completed interview, we assume we will find a larger number of interviews per day in the field and higher cooperation rates for falsifying interviewers (that is to say, a better performance). Finally, in contrast to previous research in this field, the data used here offer the possibility of including panel information from previous waves. We expect a higher probability of interview fabrication if there are (unrealistically) large deviations in a respondent's answer between two successive waves. To test this assumption, we included an indicator that determines the absolute deviation of a respondent's measured body weight (in kilograms) compared to that reported in the last interview. Table 1 presents an overview of the indicators used in our analyses as well as our assumptions regarding the occurrence of interview falsifications for each of them.

Table 1: Indicators and hypotheses for interview falsifications

Paradata		**CAPI data**	
Number of contact attempts	–	Duplicates	+
Interviewer notes	–	Straight-lining	+
Interview duration	–	'Other' answers	–
Number of asked items	–	'Code all that apply' answers	–
Number of interviews per day in field	+	Follow-up questions	–
Cooperation rate	+	Item non-response	–
Cooperation rate of partner	+	Extreme answers	–
		Primacy effect	–
		Proxy respondents	+
		Grip strength: test done	+
		Grip strength: rounding	–
		Deviation from last wave	+

Note: A minus (plus) sign besides the variable indicates that we assume less (more) of this respective indicator for fabricated interviews; for example, we expect a lower number of contact attempts.

Data

The present study uses data from the Survey of Health, Ageing and Retirement in Europe (SHARE; Börsch-Supan et al, 2013), which is a multidisciplinary cross-national panel study that has been conducted biannually since 2004. By collecting data on health, socio-economic status, and social and family networks from individuals aged 50 and older, it strongly contributes to the understanding of the ageing process in Europe.[2] Data collection is conducted face to face using a centrally developed CAPI system with an *ex ante* harmonised questionnaire for all countries. All SHARE samples are based on probability samples with full coverage of the 50+ population (for details on the national sampling frames see De Luca et al, 2015). In wave 7, data from 27 European countries (Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain (including the region of Girona), Sweden and Switzerland) plus Israel were collected. Overall, nearly 2,000 interviewers conducted over 70,000 interviews. The household response rate for refreshment samples in wave 7 varied from 40% to 63% across countries, according to standards set by the American Association for Public Opinion Research (see AAPOR, 2016: RR3). Magnitude and country differences hence are largely comparable to those in other cross-national surveys in Europe, such as the European Social Survey (ESS, 2016). Retention rates for individuals who participated in wave 6 and wave 7 varied between 62% and 92% (Bergmann et al, 2019).

Similar to other large-scale studies (Blasius and Thiessen, 2015, for example), SHARE has occasionally experienced cases of interviewer fabrication in its nearly 15 years of existence and has repeatedly removed a small number of fabricated interviews from public data releases. Among other things, the detection of these cases was possible because SHARE had implemented several quality back checks since the beginning. These include the verification of a minimum of 20% of each interviewer's complete interviews by supervisory personnel of the survey agencies in each country. However, specific (statistical) procedures for identifying falsifying interviewers and the consequences of detection were underspecified prior to wave 7. In general, survey agencies implemented some broad controls on their own to be certain that the delivered data reached an acceptable level of quality. These controls were based on a random sample of interviews that was checked by calling the selected respondents and asking them if they had participated in the interview at all and if they had answered some questions in the recorded way. Prior to wave 7, however, there were no common rules specifying what agencies should actually do or ask, and thus, there was substantial variation in the applied checking mechanisms among the participating countries in SHARE.

A case of interview fabrication in SHARE wave 6 and the development of a back-check procedure to prevent curbstoning

In this context, the survey agency of one country notified us that they had observed irregularities in their collected data after the fieldwork for wave 6 was completed. As it turned out, a regionally operating small group of interviewers delivered fabricated interviews with supposed respondents who were never truly visited. Consequently,

all cases that had been assigned to those interviewers (approximately 9% of the gross sample of drawn individuals including partners) were flagged in the data-cleaning process and excluded from any further data releases. While these suspicious interviews were concentrated in four partly adjacent regions, the representativeness of the remaining sample was not severely affected due to the large interview staff that conducted enough proper interviews in these regions. Although it was detected long before the official release of wave 6 and thus was not problematic for users of SHARE, the damage consisted of the already incurred costs, lost reputation, time-consuming data corrections and a much smaller baseline sample for analyses than had been originally planned; this damage was substantial not only for the concerned survey agency but also for SHARE.

The experience of this data fraud showed several things. First, the implemented back-check procedure in SHARE is able to detect curbstoning – at least if the concerned survey agency, as in this case, is committed to delivering high-quality data. As such, the detection of the fraud could have marked the end, as nothing actually detracted from the quality of the released SHARE data. However, we decided to adopt another strategy. In addition to being a useful learning process in several respects, the detection of a relatively large number of suspected and confirmed falsifiers (18 interviewers, accounting for 686 completed interviews including partner interviews that were declared to be falsifications based on the survey agency's suspicions) offers the possibility of exploring the underlying mechanisms of different indicators, which might help identify similar cases of future curbstoning. Such a situation is rare, especially in longitudinal survey research, because normally confirmed falsifiers are limited to only a handful of cases or intentionally manipulated experimental data (such as Menold et al, 2013). Therefore, we used the available information on verified fake and honest interviews that were carefully checked for correctness to build a model that predicts curbstoning. More precisely, we used the large set of indicators derived from paradata and CAPI data that was presented in the previous section in a multivariate cluster analysis (see, for example, Härdle and Simar, 2015; see also Kaufman and Rousseeuw, 2005 for a general introduction to cluster analysis), as this strategy proved to be superior to the application of a single indicator (see Bredl et al, 2012).

The basic idea of a cluster analysis is to group similar elements (in our case, interviews or interviewers) together, and elements in different clusters should be distant. For the assignment of interviews to one cluster, a number of different methods are available. Here, we applied the Euclidian distance as a frequently used measure to group interviews. In addition, we took into account the variations within and between interviewers over the survey's field period by standardising the indicators at the interviewer level; that is, we explored deviations in interviews conducted by a certain interviewer. The concrete indicators we used in our cluster analysis are the following (see also Table A1 in the appendix):

1 Number of contact attempts: the frequency of all contact attempts for one specific interview that were recorded by the interviewer via telephone, in person or by other means.
2 Interviewer notes: a dichotomous variable indicating whether an interviewer made at least one note (either regarding the person(s) living in the household or a specific question).

3 Interview duration: the duration of the complete interview (in minutes) based on all CAPI modules that were asked.
4 Number of asked items: the number of items that were asked by the interviewer within the entire interview.
5 Number of interviews per day in the field: the number of completed interviews conducted by an interviewer divided by the number of days passed since his/her first interview.
6 Cooperation rate/cooperation rate of partner: the number of complete interviews divided by the total number of interviews (complete plus partial) plus the number of non-interviews that include contact with an eligible respondent (refusal and break-off plus other) (see AAPOR, 2016: COOP1); the cooperation rate of partners in households with two eligible persons is a dichotomous variable indicating whether an interview with the partner was conducted.
7 Duplicates: the number of identical answers for all CAPI modules; an interview is marked as a duplicate if the questions in at least one module show the same answer pattern.
8 Straight-lining: the frequency of selecting the same answer category across all items in three multi-item sets; this value is standardised by the number of items, taking into account that identical answer patterns are more likely based on fewer items.
9 Other answers: the frequency of items across all questions for which an 'other' category is available in the questionnaire.
10 Code all that apply answers: the frequency of selecting more than one answer option based on five items for which this is possible.
11 Follow-up questions: the frequency of choosing 'no' in four filter questions with many follow-up questions.
12 Item non-response: the number of missing values across all substantial items in the presented questionnaire.
13 Extreme answers: a dichotomous variable indicating whether the (absolute) extreme values on two 11-point scales were chosen.
14 Primacy effect: the frequency of choosing the first answer category in a list of possible answer options based on four variables offering such lists.
15 Proxy respondent: a dichotomous variable indicating whether a respondent is assisted by a so-called proxy respondent in case physical and/or cognitive limitations make it too difficult for him/her to complete the interview by himself/herself.
16 Grip strength test performed: a dichotomous variable indicating whether the grip strength measurement was conducted.
17 Grip strength rounding: a dichotomous variable indicating whether multiples of five and ten were recorded by the interviewer when the test was conducted.
18 Deviation from the last wave: absolute deviation in the measured body weight (in kilograms) of the respondent between wave 6 and wave 7.

We focused on the fabrication of complete interviews as the most drastic form of interview falsification; therefore, we applied the commonly used k-means algorithm[3] to distinguish between two groups of interviewers: those that honestly interviewed their assigned respondents and hence produced valid interviews and those that fabricated their interview data (see Rokach and Maimon, 2005 for a discussion of the pros and cons of the k-means algorithm).

Table 2: Sensitivity and specificity of the identification procedure in wave 6

		True state of interview	
		Fabricated	***Valid***
State of interview according to identification procedure	***Fabricated***	90.7% (n = 622)	4.7% (n = 117)
	Valid	9.3% (n = 64)	95.3% (n = 2,370)

Data: SHARE wave 6 (end of fieldwork); only one country with confirmed information on falsifications.

The crucial question in this respect is one of sensitivity and specificity, that is, how well can we predict truly faked interviews while minimising false alarms with the indicators at hand? The answer to this question can be derived from Table 2. It shows that our model performs extremely well in wave 6. While we were able to correctly classify 91% (n = 622) of the interviews that were declared fabricated (N = 686), our model erroneously predicts only 5% (n = 117) of all interviews as fabricated when they were actually valid (N = 2,487). Obtaining a low number of these false negatives is important in our case because these interviews must be checked at great expense by the survey agency when the true state is unknown.

Implementation of the back-check procedure in SHARE wave 7

Based on these promising results, we started fieldwork preparations for the seventh wave of SHARE, beginning in February 2017 and lasting for approximately eight months. In contrast to wave 6, we now had no information on the true state of an interview, that is, if it was fabricated or not. Therefore, we had to assume a similar behaviour of data falsifiers in all participating countries and hence used the same model as in wave 6 to flag suspicious interviews and at-risk interviewers that the survey agencies then had to check. The implementation of this procedure during fieldwork was communicated to survey agencies and country teams in advance as an attempt to determine how suitable a focused sample is compared to random back checks for the identification of fabricated interviews. The decision to base our back checks on a cluster analysis and on the indicators used was shared beforehand with the survey agencies to increase their willingness to cooperate. During the fieldwork, new interview data were synchronised every two weeks. We started to run the cluster analysis as soon as at least ten interviewers were in the field and a minimum of 500 interviews were conducted in the country to ensure that our analyses are based on a sufficiently large database and thus were robust with respect to outliers. As in wave 6, all indicators were standardised at the interviewer level. In addition, we further standardised all indicators by country and questionnaire version (panel interviews vs refreshment/baseline interviews) to adequately reflect systematic differences that otherwise would bias results. We only analysed main interviews because end-of-life interviews (an interview that collects information on a deceased respondent regarding the last months of the respondent's life) with a proxy respondent (most often the partner or a close relative) are based on a very different questionnaire.

Based on these data, we applied country-specific cluster analyses distinguishing completed interviews in two cluster groups for each country that fulfilled the previously mentioned criteria: fabricated versus valid interviews. In a next step, we used this binary variable (fabricated yes/no) as the dependent variable in a logistic

regression, with all indicators (listed in the preceding section) as independent variables. The logistic regression for all available interviews is specified as follows:

$$\log(\gamma_i) = \beta_0 + \beta x_i + e_i,$$

where γ_i is the probability for an interview (i = 1,..., n) being a falsification, and β_0 is the constant term. Further, x_i denotes a vector set of explanatory variables (see the full list given earlier), while β is the corresponding vector of coefficients. Finally, e_i is the error term of the equation.

To reduce the number of possible false alarms, we only flagged the most suspicious interviews, which had a predicted probability of being fabricated in the logistic regression greater than 95%. Furthermore, an interviewer was only flagged when more than 50% of his/her completed interviews were flagged as suspicious to avoid false positives. For these cases, we sent a list of anonymised interviewer IDs to the respective survey agency and requested the agency to check at least three interviews of every listed interviewer by recontacting the respective interviewed households. If the survey agency detected any irregularities, all interviews of a certain interviewer were checked. These checks were performed mostly via telephone in addition to the classic random back checks. The results of all back checks had to be documented in a template that was provided to the survey agencies. This whole procedure was repeated roughly every four weeks during the survey's fieldwork period with the cumulative dataset. As a consequence, it was possible that the same interviewers were identified as being at risk several times.

Results

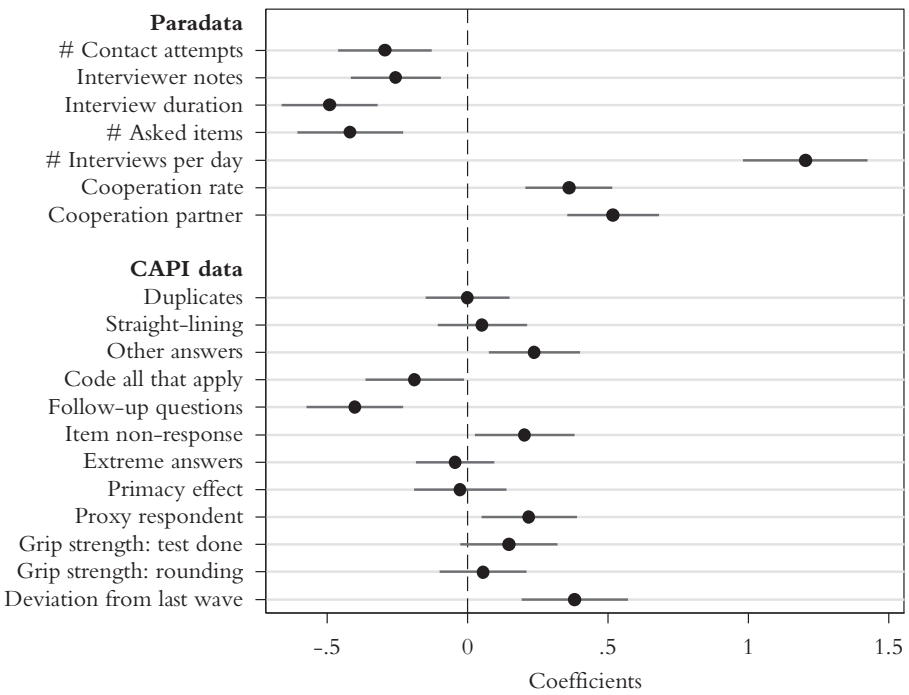
The first time we ran our identification procedure to detect fabricated interviews was the end of March 2017, approximately four weeks after the start of fieldwork in most countries. Based on a sample of 1,621 interviews from 137 active interviewers, we derived two clusters with sizes of 48.4% and 51.6%. Although most indicators suggest that the slightly smaller cluster is the predicted fabricated cluster (such as more interviews per day in the field, a shorter interview duration and fewer items asked), the nearly equal distribution of the clusters clearly illustrates the importance of reducing the suspicious cases to those that are most likely to have been fabricated. After regressing the derived cluster solution on the previously mentioned indicators, we identified 88 suspicious interviews from nine interviewers in the field as 'at risk' for curbstoning. We hence informed the concerned survey agencies and provided them with a list of the anonymised IDs of these interviewers. Based on this list, the survey agencies had to check at least three interviews of these at-risk interviewers to determine whether the initial suspicion could be verified.

After the first round of focused back checks, our suspicions were not confirmed. While this finding could be interpreted as a good sign that curbstoning did not take place, the survey agencies' back checks suffer from two problems. First, despite their effort to contact all concerned respondents – some survey agencies even conducted personal visits when a certain number of contact attempts by telephone were not successful – several back checks could still not be realised. Second, in some cases, respondents, particularly those who were very old, could not remember being interviewed, while the respective interviewer insisted on having conducted the interview. In both cases, our suspicion could not be verified, and we had to count

all interviews from the listed interviewers as valid. Approximately every four weeks, we repeated the identification procedure with the steadily increasing sample of conducted interviews and provided the survey agencies with anonymised IDs of at-risk interviewers. In this respect, the logistic regression offers the possibility to more closely investigate the patterns and the predictive power of each indicator to identify fabricated interviews.

Figure 1 shows the results of the logistic regression with the listed indicators as independent variables distinguishing between fabricated and valid interviews based on the whole interview sample of 70,133 interviews at the end of the fieldwork for wave 7 in October 2017.[4] Larger absolute coefficients indicate larger effects of the respective indicator that are significant at the 95% level when the confidence interval does not overlap with the dashed zero line. The large positive coefficient of the item 'Interviews per day', for example, indicates that a high number of interviews per day in the field is correlated with a higher probability of falsifying interviews. Hence, indicators derived from paradata seem to work better for distinguishing between the two clusters than indicators from the CAPI interview that show, on average, more overlap with the zero line. Apart from being highly significant in differentiating between fabricated and valid interviews, the indicators based on paradata also confirm our hypotheses. Interviewers stating that they contact their assigned respondents less often, make fewer notes and have shorter interviews with fewer asked items are more likely to deliver fabricated interviews.

Figure 1: Predictive power of used indicators to distinguish between fabricated and valid interviews



Note: Logistic regression coefficients with 95%-confidence intervals; data: SHARE wave 7 (end of fieldwork).

Moreover, those interviewers show a much better performance, which is reflected in the average number of interviews conducted per day in the field, and achieve higher cooperation rates, both for the assigned respondents and their partners. This is an interesting aspect because it shows that a very good performance (sometimes too good to be true) in terms of cooperation/response rates should be carefully evaluated. Additionally, this clearly holds some conflict potential because high response rates are in the interest of both the survey agencies and the public or scientific institutions that run the surveys.

In contrast, only some indicators that are directly derived from the CAPI interview can help distinguish between fabricated and valid interviews. In particular, this holds true for follow-up questions (more follow-up questions are correlated with a lower probability of curbstoning) and deviations in recorded answers compared to the last wave (more deviations are related to a higher probability of curbstoning). Both findings support our hypotheses. In particular, the last indicator using panel information from previous waves shows the potential of applying our identification procedure in longitudinal surveys. Our hypotheses for 'code all that apply' answers and the use of proxy respondents were confirmed to a lesser degree. Two other indicators – the frequency of the selection of the residual category 'other' in questions with several response options and the number of missing values for sensitive questions – although slightly significant do not support our hypotheses. The other indicators, while largely in line with our hypotheses, do not reach the significance level of 95%.

In addition, Table 3 shows the result of the cluster analysis at the end of fieldwork in wave 7. Based on our identification procedure, the survey agencies checked 1,226 suspicious interviews from at-risk interviewers out of 70,133 interviews overall by recontacting the concerned households. Among those, two flagged interviewers could be convicted of curbstoning. Overall, 52 interviews from these interviewers could be verified as having been fabricated. The random back checks on the other hand identified four interviewers with 67 fabricated interviews out of a total of 28,719 checked cases. Therefore, although our model has a rather low sensitivity (43.7%), that is, we only identified approximately half of all fabricated interviews that could be verified by the survey agencies, our targeted back-check procedure seems to be more efficient than the random back checks when taking into account the number of interviews that actually have been checked based on the respective procedure. The rather low sensitivity, of course, is partly due to our conservative approach of only flagging the most suspicious interviews with the aim of reducing the survey agencies' effort to check valid interviews with a lower probability of having been fabricated. Hence, the rate of false positives, that is, interviews that turned out to be valid after having been flagged as being suspicious based on our identification procedure, is very low (only approximately 2%). However, it must be noted that the number of fabricated interviews that have been verified by the

Table 3: Sensitivity and specificity of the identification procedure in wave 7

		True state of interview	
		Fabricated	***Valid***
State of interview according to identification procedure	***Fabricated***	43.7% (n = 52)	1.7% (n = 1,174)
	Valid	56.3% (n = 67)	98.3% (n = 68,840)

Data: SHARE wave 7 (end of fieldwork).

survey agencies is also very low: only 119 interviews from six interviewers could be verified as curbstoning.

Conclusion

Curbstoning, the fabrication of an entire interview, is a rare event in SHARE but can nevertheless lead to negative consequences regarding the panel sample, such as a loss in sample size or the need for time-consuming data corrections of information collected during previous waves. Consequently, we developed a technical procedure to identify interview fabrication and deal with it during ongoing fieldwork in the seventh wave of SHARE, rather than waiting until the data collection has been completed. Overall, we can summarise that our identification procedure based on a multivariate cluster analysis is able to identify fabricated interviews, but additional random back checks are useful to increase the number of detected curbstoning cases. This finding confirms previous research (Bredl et al, 2013, for example) and provides further evidence that neither focused nor random back checks alone are sufficient for identifying fabricated interviews. In addition, we found that paradata in our case works better than CAPI data for predicting interview fabrication. The variables that perform best are mostly performance indicators: at-risk interviewers show a significantly better (perhaps too good) performance in terms of realising cooperation with both an assigned respondent and a possible partner and conducting more interviews in a shorter time period. As interviewers are paid per completed interview in SHARE, the incentive structure seems rather straightforward. In line with that, shorter interviews with fewer items asked (probably due to the avoidance of time-consuming follow-up questions) are further powerful predictors for suspicious – and ultimately fabricated – interviews. In addition, the use of panel information from previous waves also significantly helps distinguish between fabricated and valid interviews. This is an important finding that should be focused on more in future research using panel data.

In addition to these results, we must state that our identification procedure worked better when the true outcome (that is, whether the interviews were fabricated) was known. It did not perform quite as well in the case in which our goal was to equip survey agencies with a focused sample of at-risk interviewers, which then needed to be checked to confirm our initial suspicion. In our opinion, there are several reasons for this. First, the small amount of (detected) fabrications makes it more difficult for any statistical identification procedure to identify curbstoning. The number of verified interview fabrications in our case was not large enough. This, of course, is good news, as one could argue that data quality in SHARE is not severely affected by curbstoning. Although not clearly verifiable, we believe that the mere announcement of detailed interview back checks, both to survey agencies and to interviewers during the training sessions before starting fieldwork, has contributed to this result. This is not to say that we can be perfectly sure that curbstoning is not an issue at all in SHARE, but at least there are no obvious signs of large-scale interview fabrications. Second, SHARE is a cross-national survey that covers very different countries from Finland to Greece and Portugal to Romania. Thus, it might be possible that interviewers (but also survey agencies; see Blasius and Thiessen, 2018) behave differently in different contexts. We tried to take this into account by standardising the indicators used by country and by applying country-specific cluster analyses. However, there is still the possibility that indicators work differently or follow diverse mechanisms

in different countries. Thus, using a pooled logistic regression model to identify the most suspicious interviews might have resulted in a more vague classification that presumably prevented a clear identification of fabricated interviews. In this respect, there is a third important aspect that should be considered. As already explained, we followed a very conservative approach and only selected the most suspicious cases for back checks by the survey agencies. Further investigations showed that those fabricated interviews, which were not detected by our cluster analysis but by the additional random back checks of the survey agencies, exhibit a probability slightly below our cut-off criteria of 95%. Thus, it might be helpful to reduce the threshold, although this means a higher chance of false negatives and thus higher costs. We are not aware of other studies that have published detailed information on that. Therefore, we hope to foster further research and explorations in this direction.

While all these aspects refer more or less to the statistical implementation of the identification procedure, the most important question in our view is how to precisely confirm initial suspicions. Our example clearly shows that giving survey agencies a list of interviewers for back checking is not enough to be sure if a certain interview has been fabricated. What seems clear is that survey agencies play a key role in this respect, as they – at least in SHARE – are the only ones that can legally contact their interviewers. Most important, therefore, is a close collaboration among all involved partners (survey agencies including interviewers on the one side and the scientific institution responsible for the survey on the other side) and a sincere commitment regarding the provision of the highest possible data quality, including the disclosure of falsifications. Moreover, a comprehensive concept with respect to back checks is needed that is ideally developed jointly and clearly outlines who and what should be checked and how this has to be documented. Therefore, the statistical procedure to identify fabricated interviews must be embedded in a broader framework of data quality monitoring that is not only focused on the detection of curbstoning but also considers its prevention (such as careful interviewer training including various feedback loops) and describes strategies that should be employed with interviews and interviewers after the detection of curbstoning. All this shows that there are still many open questions. Nevertheless, we hope that the present study fosters further research in this area, which is definitely needed but far too often is not being published for fear of negative consequences with respect to reputation and funding. Therefore, we want to suggest a more open handling of issues such as interview fabrication because a strong commitment to solid scientific standards is needed in order to both create distance from purely marketing studies with sometimes dubious business practices and convince the public that participating in a study such as SHARE has social value.

Acknowledgments

This paper uses data from SHARE wave 6 and wave 7 (w6_internal_release_ivcheck, w7_internal_release_ivcheck); see Börsch-Supan et al (2013) for methodological details. 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.

Notes

[1] Corresponding author.

[2] In addition, SHARE is harmonised with similar panel surveys in the British Isles, the United States, Japan, Korea, China, India, Mexico, Brazil and South Africa.

[3] All our analyses were performed with Stata 14. For the cluster analysis, we used the *kmeans* command; regression analyses were based on the *logistic* command. All scripts are available on request from the authors.

[4] The results were consistent when we ran the cluster analysis repeatedly during fieldwork.

Conflict of interest

The authors declare that there is no conflict of interest.

References

AAPOR (American Association for Public Opinion Research) (2016) *Standard definitions: Final dispositions of case codes and outcome rates for surveys*, Oakbrook Terrace, IL: AAPOR.

Allen, M. (2017) *The SAGE encyclopedia of communication research methods*, Thousand Oaks: Sage.

Bergmann, M., Kneip, T., De Luca, G. and Scherpenzeel, A. (2019) *Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), wave 1–7*, SHARE Working Paper Series 41-2019, Munich: SHARE-ERIC.

Blasius, J. and Thiessen, V. (2013) Detecting poorly conducted interviews, In P. Winker, N. Menold and R. Porst (eds), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention*, Frankfurt am Main: Peter Lang, pp. 67–88.

Blasius, J. and Thiessen, V. (2015) Should we trust survey data? Assessing response simplification and data fabrication, *Social Science Research*, 52: 479–93. doi: 10.1016/j.ssresearch.2015.03.006

Blasius, J. and Thiessen, V. (2018) Perceived corruption, trust, and interviewer behavior in 26 European countries, *Sociological Methods & Research*, Online First: https://doi.org/10.1177/0049124118782554

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S. and Zuber, S. (2013) Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE), *International Journal of Epidemiology*, 42(4): 992–1001.

Bredl, S., Storfinger, N. and Menold, N. (2013) A literature review of methods to detect fabricated survey data, In P. Winker, N. Menold and R. Porst (eds), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention*, Frankfurt am Main: Peter Lang, pp. 3–24.

Bredl, S., Winker, P. and Kötschau, K. (2012) A statistical approach to detect interviewer falsification of survey data, *Survey Methodology*, 38(1): 1–10.

Bushery, J.M., Reichert, J.W., Albright, K.A. and Rossiter, J.C. (1999) *Using date and time stamps to detect interviewer falsification*, JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, pp. 316–20.

De Luca, G., Rossetti, C. and Malter, F. (2015) Sample design and weighting strategies in SHARE wave 5, In F. Malter and A. Börsch-Supan (eds), *SHARE wave 5: Innovations and methodology*, Munich: MEA, Max Planck Institute for Social Law and Social Policy, pp. 75–84.

Ericksen, E.P. and Kadane, J.B. (1985) Estimating the population in a census year: 1980 and beyond – rejoinder, *Journal of the American Statistical Association*, 80(389): 98–109.
ESS (2016) *ESS7: 2014 documentation report*, London: ESS ERIC.
Härdle, W.K. and Simar, L. (2015) *Applied multivariate statistical analysis* (4th edn), Heidelberg: Springer.
Hood, C.C. and Bushery, J.M. (1997) *Getting more bang from the reinterviewer buck: Identifying 'at risk' interviewers*, JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association, pp. 820–24.
Japec, L. (2006) Quality issues in interview surveys: some contributions, *Bulletin of Sociological Methodology*, 90(1): 26–42. doi: 10.1177/075910630609000104
Kaminska, O., McCutcheon, A.L. and Billiet, J. (2010) Satisficing among reluctant respondents in a cross-national context, *Public Opinion Quarterly*, 74(5): 956–84. doi: 10.1093/poq/nfq062
Kaufman, L. and Rousseeuw, P.J. (2005) *Finding groups in data: An introduction to cluster analysis*, Hoboken, NJ: John Wiley & Sons.
Koch, A. (1995) Gefälschte interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994, *ZUMA-Nachrichten*, 19(36): 89–105.
Koczela, S., Furlong, C., McCarthy, J. and Mushtaq, A. (2015) Curbstoning and beyond: Confronting data fabrication in survey research, *Statistical Journal of the IAOS*, 31(3): 413–22. doi: 10.3233/SJI-150917
Krosnick, J.A. and Alwin, D.F. (1987) An evaluation of a cognitive theory of response-order effects in survey measurement, *Public Opinion Quarterly*, 51(2): 201–19. doi: 10.1086/269029
Kuriakose, N. and Robbins, M. (2016) Don't get duped: Fraud through duplication in public opinion surveys, *Statistical Journal of the IAOS*, 32(3): 283–91. doi: 10.3233/SJI-160978
Li, J., Brick, J.M., Tran, B. and Singer, P. (2011) Using statistical models for sample design of a reinterview program, *Journal of Official Statistics*, 27(3): 433–50.
Menold, N., Winker, P., Storfinger, N. and Kemper, C.J. (2013) A method for ex-post identification of falsifications in survey data, In P. Winker, N. Menold and R. Porst (eds), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention*, Frankfurt am Main: Peter Lang, pp. 25–47.
Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. and Kennet, J. (2004) *A system for detecting interviewer falsification*, JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association, pp. 4968–75.
Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O. and Hsieh, Y.P. (2016) Interviewer falsification: current and best practices for prevention, detection, and mitigation, *Statistical Journal of the IAOS*, 32(3): 313–26. doi: 10.3233/SJI-161014
Rokach, L. and Maimon, O. (2005) Clustering methods, In O. Maimon and L. Rokach (eds), *Data mining and knowledge discovery handbook*, Boston, MA: Springer, pp. 321–52.
Schäfer, C., Schräpler, J.-P., Müller, K.-R. and Wagner, G.G. (2005) Automatic identification of faked and fraudulent interviews in the German SOEP, *Schmollers Jahrbuch: Journal of Applied Social Science Studies/Zeitschrift für Wirtschafts-und Sozialwissenschaften*, 125(1): 183–93.
Schräpler, J.-P. (2004) Respondent behavior in panel studies: a case study for income nonresponse by means of the German Socio-Economic Panel (SOEP), *Sociological Methods & Research*, 33(1): 118–56.

Schräpler, J.-P. (2010) *Benford's law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP)*, SOEP Papers on Multidisciplinary Panel Data Research, No. 273, Berlin: DIW.

Schräpler, J.-P. and Wagner, G.G. (2003) *Identification, characteristics and impact of faked interviews in surveys: An analysis by means of genuine fakes in the raw data of SOEP*, IZA Discussion Paper No. 969, Berlin: DIW.

Schreiner, I., Pennie, K. and Newbrough, J. (1988) *Interviewer falsification in census bureau surveys*, JSM Proceedings, Survey Research Methods Section, Alexandria: VAL American Statistical Association, pp 491–6.

Schupp, J. (2018) *Umfragebasierte Studien: 'Fake-Interviews' bleiben die Ausnahme*, DIW Wochenbericht No. 6, Berlin: DIW.

Shaeffer, E.M., Krosnick, J.A., Langer, G.E. and Merkle, D.M. (2005) Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions, *Public Opinion Quarterly*, 69(3): 417–28. doi: 10.1093/poq/nfi028

Turner, C.F., Gribbe, J.N., Al-Tayyib, A.A. and Chromy, J.R. (2002) *Falsification in epidemiologic surveys: Detection and remediation*, Technical Papers on Health and Behavior Measurement, No. 53. Washington, DC: Research Triangle Institute.

Werker, H.F. (1981) Results of the 1980 US census challenged, *Population and Development Review*, 7(1): 155–67. doi: 10.2307/1972793

Yamamoto, K. and Lennon, M.L. (2018) Understanding and detecting data fabrication in large-scale assessments, *Quality Assurance in Education*, 26(2): 196–212. doi: 10.1108/QAE-07-2017-0038

Appendix

Table A1: Description of indicators

Variable	Min	Max	Mean	SD
Number of contact attempts	0	23	2.12	1.40
Interviewer notes	0	1	.18	.23
Interview duration	3.09	163.64	57.29	16.12
Number of asked items	43	972	324.51	43.12
Number of interviews per day in the field	.01	7	.58	.42
Cooperation rate	.05	1	.76	.15
Cooperation rate of partner	0	1	.62	.15
Duplicates	0	5	2.58	.42
Straight-lining	0	.95	.48	.06
Other answers	0	2.14	.52	.26
Code all that apply answers	0	1.75	.93	.30
Follow-up questions	.65	3	1.35	.17
Item non-response	0	34	1.38	1.79
Extreme answers	0	1	.16	.13
Primacy effect	0	2.5	.50	.23
Proxy respondent	0	1	.05	.07
Grip strength: test done	0	1	.90	.12
Grip strength: rounding	0	1	.25	.11
Deviation from last wave	0	25	1.44	1.66

Data: SHARE wave 7 (end of fieldwork, N = 70,133); SD = standard deviation.