SCHRIFTEN ZUR
EMPIRISCHEN WIRTSCHAFTSFORSCHUNG
Herausgegeben von Peter M. Schulze und Peter Winker
22

Peter Winker / Natalja Menold / Rolf Porst
(eds.)

Interviewers' Deviations in Surveys

Impact, Reasons, Detection and Prevention

PL ACADEMIC RESEARCH

A Literature Review of Methods to Detect Fabricated Survey Data

Sebastian Bredl, Nina Storfinger, Natalja Menold

Abstract

This paper reviews literature dealing with the issue of detecting interviewers who falsify survey data. The most reliable method of detecting falsifiers is through face-to-face reinterviewing of survey participants. However, especially in large scale surveys only a limited number of participants can usually be reinterviewed. A review of the present literature clearly indicates that reinterviewing is more effective if the re-interview sample is based on some indicators that might comprise metadata, survey data, or interviewer characteristics. We examine relevant literature with regard to the suitability of different types of indicators that have been used in this context.

Keywords: Interviewer falsification, quality control of survey data, reinterview

Acknowledgements

Financial support through the DFG in project WI 2024/2-1; ME 3538/2-1 within SPP 1292 is gratefully acknowledged. Furthermore we would like to thank Andreas Diekmann, Gesine Güllner and Peter Winker for their valuable comments on previous versions of the paper.

Introduction

In economic and social research, survey data is often the cornerstone of empirical investigations. Several factors that may impair the quality of such data during the period of field work, such as systematic non-response or interviewer effects on response behaviour, have gained attention in literature. Another important factor that has not received as much attention thus far is the conscious deviation from prescribed procedures by the interviewer, which is referred to as interviewer falsification (Schreiner et al., 1988) or cheating (Schräpler and Wagner, 2003). The American Association for Public Opinion Research (AAPOR) defines behaviour like this as 'intentional departure from the designed interviewer guidelines and instructions, unreported by the interviewer, which could result in the contamination of data' as 'interviewers' falsification' (2003: 1). There is a wide range of potential forms of cheating (cf. also Schräpler, 2010). The most blatant of these is undoubtedly the fabrication of entire interviews without ever having contacted the target person. Another possibility is partial fabrication, for example by making the contact but only asking some of the questions contained in the questionnaire and faking the remaining data (Harrisson, 1947). More subtle forms are listed by Case (1971), who mentions inter-

viewing someone other than the intended person, changing the interview mode, or changing the location of the interview. The present chapter reviews literature dealing with detecting fabrication of complete interviews as the most blatant form of cheating as well as literature dealing with the detection of partial falsification.

Seen from the interviewer's perspective, there are several reasons why data fabrication might be an attractive option. Interviewers do not usually have a strong interest in delivering high-quality data, apart from the potentially satisfying feeling of having done a good job. As Durant (1946: 290) puts it, '[o]ne day's interviewing, however well done, merely serves to lead on to the next day's interviewing'. Furthermore, interviewers have to ask people who they do not know to reveal personal information, which may trigger dismissive reactions (cf. Crespi, 1945, Stewart and Flowerman, 1951, Köhne-Finster and Güllner, 2009) and are often faced with payment schemes based largely on the number of completed interviews (Kennickell, 2002). This might create pressure to augment 'quantity' and neglect the 'quality' of interviews, and may finally promote conditions leading to data fabrication (cf. Bennett, 1948, Sudman, 1966).

So far very little research has been done on the consequences of data fabrication for subsequent statistical analyses. This might be due in part to the fact that the severity of these consequences is obviously related to the prevalence of data fabrication. This parameter can be estimated only roughly, as it is likely that not all relevant cases can be detected. Studies reporting some estimates (e.g. Schreiner et al., 1988, Koch, 1995, Krejsa et al., 1999, Schräpler and Wagner, 2005, Li et al., 2009) suggest that the proportion of fabricated interviews rarely exceeds 5%. However, these studies refer only to large-scale surveys. In smaller surveys, with only a handful of interviewers, one may observe much larger proportions of fabricated interviews (Harrison and Krauss, 2002, Bredl et al, 2012). Not only is the quantity of fabricated data an important determinant in this context, but so is quality. If cheaters were able to reproduce "realistic" data, there would hardly be a problem. According to several studies (Hippler, 1979, Reuband, 1990, Schnell, 1991, Schräpler and Wagner, 2005), cheaters generally do quite a good job of fitting their data to marginal distributions found in real data, but they struggle to reproduce more complex relationships like those revealed by factor analysis or multivariate regression analyses. Consequently, even a small proportion of fabricated interviews, say of around five percent, might have a severe impact on the results of multivariate statistical analysis as shown by Schräpler and Wagner (2005). But this is not necessarily the case as demonstrated by Schnell (1991).

As interviewer data fabrication seems to be a non-negligible problem, one must be concerned about how to detect fraudulent interviews. Although the

overall volume of literature on this issue is still modest, the variety of proposed methods and indicators is quite considerable, which clearly calls for some comparison and evaluation of different approaches. This is the issue we would like to address in this literature review. Based on our analyses we also try to formulate some recommendations on how to proceed in order to detect fabricated data, and we identify fields of research that need more attention in the future.

For our literature review we systematically searched different data bases for the social and economic sciences. Thereby, we analysed literature, published in English and German. Of the literature found dealing with complete or partial fabrication of interviews, the majority concerned methods of detecting falsifiers (most were journal articles, but conference proceedings and working papers were also available). In our review, we considered contributions on methods of detection based on empirical data. Overall, our search results show that up to now no extended research exists on the topic of falsifications. Nevertheless, we were able to find interesting results with respect to detection methods and to discuss the advantages and disadvantages of the different methods.

In the section "Overview of Key Studies", we examine five key studies which either applied detection methods during the field control in order to identify falsifiers (ex-ante studies) or tested the performance of several methods using datasets with known cases of falsification (ex-post studies). The aim of ex-post studies is to identify indicators that differ for data collected honestly and data which has been falsified. Based on this examination, the section "Overview of Different Approaches" discusses different methods for detecting data fabrication. Here, we focus on the effectiveness and the generalisability of the respective method. "Discussion and Outlook" summarizes the findings of our literature review and formulates some recommendations based on insights from the previous sections. Furthermore, this section highlights fields in which more research is needed.

Overview of Key Studies

In this section, we characterise selected comprehensive studies dealing with the detection of fabricated data. Table 1.1 provides an overview of these studies. As mentioned above, we distinguish between ex-ante studies employing the respective methods in order to detect falsifiers and ex-post studies that tested several indicators in datasets with known cases of data fabrication. All ex-ante studies included in the table used recontact procedures combined with other methods. With respect to the proportion of fabricated interviews we provide two numbers for ex-ante studies: the first refers to the proportion of falsified interviews in a

random recontact sample, the second to the proportion obtained when recontact procedures were combined with other methods. Within ex-ante and ex-post studies different data analyses were conducted, using meta-data or collected survey data. Metadata, also called para-data, are survey process data, such as contact outcomes, obtained by interviewers or data produced during the interview (e.g. with the help of time stamps). Other analyses of survey data include a comparison of answers to survey questions, response sets (or response behaviour), and the application of Benford's Law.

Table 1.1: Selected studies dealing with the detection of data fabrication

Authors	**Survey**	**Share fabricated Interviews**	**Detection methods:** Recontact	Metadata	Benford's Law	Other Analyses
		Ex-ante studies				
Koch (1995)	Large scale survey; German population; ALLBUS	random: 0.4%, combined: 2.3%	**X**			**X**
Hood, Bushery (1997)	Large scale survey; US poulation; NHIS	random: 0.2%, combined: 3.6%	**X**	**X**		**X**
Turner et al. (2002)	Large scale survey; Baltimore population	49% of 451 interviews contributed by 6 falsifiers (in total: 1200 interviews)	**X**	**X**		
		Ex-post studies				
Murphy et al. (2004)	Large scale survey; US population; NSDUH	19.5% in one highly affected US-state, no information on other states		**X**		**X**
Schraepler and Wagner (2005)	Large scale survey; German population; GSEOP	Sample A:0.6%; Sample B: 1.5% Sample C: 2%			**X**	**X**

Koch (1995)

Koch (1995) describes control procedures and their results in a survey of the German population (ALLBUS, German General Social Survey, 1994). In 1994 personal registers from registration offices started being used in the ALLBUS as sample frame. The previous sample method was 'random route' (ADM-System, Heyde and Loeffler (1993)), in which interviewers selected sample units within the two last stages of the selection process.

In contrast to ADM-samples, selected persons in personal register samples were known prior to data collection. Additionally, information about gender and age of sampled persons was provided in the sample frame. Interviewers received names and addresses of selected persons and should have interviewed exactly these persons. Hence, in the ALLBUS 1994 Koch (1995) was able to systematically check for falsifications by comparing the information on gender and age in the survey data with the data from the registration offices. Overall, the control procedures combined different steps:

A portion of interviews (25%) was routinely controlled by the survey institute responsible for data collection using postcards – they obtained a 60% response rate. These controls found 15 cases which were conducted incorrectly. Hence, these controls did not reveal considerable information about problems with the data.

In addition, all 3505 interviews realised in the ALLBUS 1994 were controlled by Koch, comparing gender and age of selected and interviewed persons. All cases with deviations detected by Koch (n = 196) were controlled by a new contact (in person, by phone or by post). Fraudulent interviewer behaviour could be detected in 81 cases (2.3%), of which 45 were complete falsifications of the interview. Koch emphasizes that the detection method he used in the ALLBUS is restricted by the sample method used. Samples, which use one or more selection stages, in which interviewers are involved (random route or samples with address registers as sample frame), cannot effectively apply this method, since the selected person is – as a rule – unknown prior to data collection. Another restriction of this method is that age and gender provide insufficient information to effectively expose falsified interviews. In most cases gender is easy to determine by the target person's first name, and age could be estimated by interviewers or asked in a short interview with the target person or with other household members (even with neighbours). The use of age and gender as information can allow only for the detection of significant carelessness in interviewers' work or other technical problems in the field, for example. It seems plausible to assume that falsifiers who are more cautious are not detected by the procedure described by Koch. Thus, the level of 2.3% of detected falsifications represents a lower

bound for very crude fabrications. Nevertheless, Koch's work indicates that a more focused recontact procedure is more effective than controls conducted by the survey institute with a portion of interviewed persons who are selected without deliberate considerations.

Hood and Bushery (1997)

Hood and Bushery (1997) investigated the usefulness of several indicators in order to create a focused reinterview sample that could be applied to the US-National Health Interview Survey (NHIS). According to the authors data fabrication occurs rarely in the NHIS. As a result, many reinterviews are required to detect a falsifier. In this context the authors emphasize the usefulness of a focused reinterview that concentrates on interviewers who seem to be more likely than others to have fabricated data according to some indicators.

Hood and Bushery assume that cheating interviewers try to 'keep it simple' (p. 820). Thus, they can be expected to label eligible households as ineligible and choose answers that allow questions to be skipped, leading to avoidance of subsequent optional parts of the questionnaire. For example, a considerable number of questions was not asked in white households in the NHIS. Consequently, a high proportion of white or ineligible households within an interviewer's assignment may be a sign of data fabrication.

The basic idea behind the approach is to examine data in questionnaires as well as some metadata (ineligible households) in order to identify interviewers who merit a closer look during the reinterview stage. However, it is clear that a relatively high proportion of white or ineligible households in one interviewer's assignments is not necessarily linked to dishonest behaviour, but rather might also be due to the specific characteristics of the area were the interviews were conducted. This is known as so-called spatial homogeneity (cluster related design effect; cf. Groves et al. (2004)), meaning in this case the homogeneity of individuals living within a geographical area. To differentiate between interviewer effects and spatial homogeneity, Hood and Bushery considered the differences between actual proportions and those that could be expected based on data from the 1990 census. If differences for all variables exceeded a certain threshold, the interviewer was flagged as an outlier and was then checked using focused reinterviews.

During the focused reinterview 3 falsifiers were detected from the 83 interviewers that were checked (3.6%). This 'success rate' is clearly above the 0.2% achieved by random reinterview. Although the informative value of these numbers should not be overrated, as they rely on a small number of cases, they do

indicate that focused reinterviews deliver better results than purely random reinterviews.

The general problem with this approach is that discriminating between effects caused by data fabrication and those caused by the particularities of an interviewer's assignment is difficult. A reliable reference survey – like the 1990 census in the case of the Hood and Bushery study – is often simply not available. Furthermore – and a point also made by Hood and Bushery (1997) – in contrast to the study by Koch (1995) the approach considers interviewers and not interviewed individuals. This may be problematic if an interviewer fabricates only a small part of his assignments. In this case, indicators based on all interviews done by an interviewer might have only little discriminatory power.

Turner et al. (2002)

Turner et al. (2002) describe their painful experiences with falsifications of a large part of the sample in a Baltimore population survey. In contrast to national large scale surveys described above, this particular survey had two special aspects: firstly, it was related to a quite sensitive topic (sexually transmitted diseases) in which biological specimens were collected; secondly, it was a large local survey. This survey differs from national surveys for the second reason, since the latter does not need a large interview staff in a local area. It was particularly difficult for the data collection institute to recruit a sufficient number of interviewers in Baltimore. Turner et al. (2002) report that very low participation rates were obtained, and as a result additional interviewer trainings were conducted and the data collection period was extended.

The research team found irregularities in the data delivered by the data collection institute: six interviewers showed implausible success rates in conducting interviews. In fact 54% to 85% of assigned households were successfully interviewed by these interviewers, in contrast to other interviewers, who succeeded only in 31% of the cases on average. All interviews submitted by these six interviewers were verified by telephone or face-to-face recontact. In addition, controls for other interviewers were conducted. Here, the authors used metadata (cf. Table 1.1) to find suspect cases and combined them with a reinterview for verification. As a result it was found that 49% of the 451 interviews submitted by six suspected interviewers were falsifications.

The procedure by Turner et al. (2002) is similar to that reported by Koch (1995): research staff conducted controls independent of any controls conducted by the data collection institute. In contrast to Koch (1995), who checked only suspect cases, all interviews conducted by suspicious interviewers were controlled by Turner et al. (2002), with a high hit ratio for fabricated interviews. But

in comparison to other studies, using the number of conducted interviews as a kind of metadata is restricted by the specifics of the survey. These specifics are associated with difficulties in conducting a local population survey on a sensitive topic. However, studies we discuss in this section show that local population surveys on sensitive topics are particularly prone to falsifications, and that it would be more effective to recontact all cases assigned to a dishonest interviewer.

Murphy et al. (2004)

Murphy et al. (2004) analysed data produced by three identified falsifiers in the American National Drug Survey on Drug Use and Health (NSDUH). This large scale survey selects around 70,000 persons each year who are interviewed using computer-assisted interviewing (CAPI) and audio computer-assisted self interviewing (ACASI), in which the laptop is given over to the respondent. Hence, the laptop registered time stamps for each question and each interview step in both modes, which allowed for the calculation of elapsed time for each respective action.

Like Turner et al. (2002) Murphy et al. (2004) examined response patterns to sensitive questions related to the lifetime use of cigarettes, alcohol, marijuana, cocaine and heroin. The authors calculated the proportion of respondents per interviewer who claimed to have already consumed the respective drug during their lifetime. To account for spatial homogeneity the authors controlled for demographic characteristics of the (alleged) respondents by examining shares separately for men and women, younger and older respondents and Hispanics and non-Hispanics. The resulting indicator performed extremely well in separating falsifiers and honest interviewers. In both cases, all three falsifiers were among the top four interviewers, if interviewers were ranked according to the values of an index indicating deviations between drug abuse rates in the interviewer's sample and the remaining data. As in the study by Turner et al. (2002) it turned out that falsifiers struggle to adequately reproduce answers to very sensitive questions.

Murphy et al. (2004) employed metadata – namely time stamps – in order to determine whether response times are different when falsifiers fabricate data as compared to situations in which the data is collected honestly. The NSDUH is a very interesting application in this regard, as it consists of the CAPI and the ACASI part. However, it turned out that clear patterns of differences between falsifiers and honest interviewers did not emerge for either the CAPI part or for the ACASI part. One of the falsifiers was generally much faster than the other interviewers, but the other two falsifiers were much slower.