Zentrum für internationale Entwicklungs- und Umweltforschung der Justus-Liebig-Universität Gießen

A Literature Review of Methods to Detect Fabricated Survey Data

by

SEBASTIAN BREDL*, NINA STORFINGER** and NATALJA MENOLD***

No. 56

Gießen, August 2011

*Lehrstuhl für Statistik und Ökonometrie, Fachbereich Wirtschaftswissenschaften
Justus–Liebig–Universität Gießen
Licher Str. 64
35394 Gießen
E-Mail: Sebastian.Bredl@wirtschaft.uni-giessen.de.

**Center for International Development and Environmental Research (ZEU),
Section 3
Justus-Liebig-Universität Gießen
Senckenbergstr.3
35390 Gießen
Email: Nina.Storfinger@zeu.uni-giessen.de

***GESIS Leibniz-Institut für Sozialwissenschaften
P.O. Box 122155
68159 Mannheim
E-Mail: natalja.menold@gesis.org

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, 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 reinterview sample is put together according to some indicators that might be based on metadata, survey data or interviewer characteristics. We examine existing literature with regard to the suitability of different types of indicators that have been used in this context.

JEL: C83 C93
Keywords: Interviewer falsification, quality control of survey data, reinterview

Acknowledgements
Financial support through the DFG in project WI 2024/2-1 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.

1 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 collection stage, such as systematic non-response or interviewer effects on response behaviour, have gained attention in literature. Another important factor that has not received 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). In relation to this the American Association for Public Opinion Research (AAPOR) defines "intentional departure from the designed interviewer guidelines and instructions, unreported by the interviewer, which could result in the contamination of data" as "interviewers' falsifications"(AAPOR 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 falsification, for example making the contact but only asking a portion of the questions contained in the questionnaire and faking the remaining data (Harrisson 1947). More subtle forms are listed by Case (1971). They include interviewing someone other than the intended person (for example another family member or a neighbour), changing interview mode (for example conducting the interview by telephone when a face-to-face interview is required), or changing the location of the interview (for example conducting it on a street corner if at-home interviews are required). This article reviews literature dealing with detecting the most blatant form of cheating, namely the partial or complete fabrication of questionnaires by the interviewer.

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. Interviewers are not involved in planning surveys or developing questionnaires and it is unlikely that interviewers are trained in scientific research ethics (AAPOR 2003). Furthermore, interviewers are not involved in processing data subsequent to data collection during the field work period. As Durant (1946: p. 290) puts it, "[o]ne day's interviewing, however well done, merely serves to lead on to the next day's interviewing". Thus, the reward from doing good work might be very small, whereas the task itself can sometimes be quite unpleasant. Interviewers are required 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). Additionally, interviewers are often faced with payment schemes based largely on the number of completed interviews (Kennickell 2002), which 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) report on a survey in which only two of five interviewers delivered reliable data. Bredl et al. (2008) mention a case in which the first round of a survey conducted by four interviewers consisted entirely of faked interviews.

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 analysis. 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). To the best of our knowledge, no study has yet been published that investigates the impact of extremely high proportions of faked interviews, comparable to those that have been observed in some small-scale surveys. Considering that small scale surveys play an important role in the social sciences this topic merits attention.

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 analysis 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. Thus, we analysed literature, published in English and German. Of the literature found, 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 articles 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.

We distinguish between two types of studies. On the one hand there are articles dealing with detection methods applied during field control procedures

in surveys in order to identify unknown falsifiers. We refer to this type of study as "ex-ante studies". On the other hand we discuss studies applying different indicators to datasets with known cases of falsification. We label this type of study "ex-post study". The aim of ex-post studies is to identify indicators that differ for data collected honestly and data which has been falsified. In Section 2 we examine key studies for both types of approaches. Based on this examination, Section 3 discusses different methods to detecting data fabrication. Here we focus on the effectiveness and the generalisability of the respective method. Section 4 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.

2 Overview of key studies

In this section, we characterise selected comprehensive studies dealing with the detection of fabricated data. Table 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. The table consists of the following information: the survey to which the respective study refers, the proportion of fabricated interviews found in each of the studies, and the methods used to detect falsifiers. 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 comparison of answers to survey questions, response sets (or response behaviour), and the application of Benford's Law.

Koch (1995)

Koch (1995) describes control procedures and their results in a survey of the German population (ALLBUS, German General Social Survey, 1994). The ALLBUS is a biannual survey related to social issues which was first conducted in 1985. The study by Koch was motivated by a discussion about falsified interviews in the German press. This discussion revealed the impression that data arising from face-to-face surveys are, as a rule, contaminated by falsifiers. To counter this impression Koch published the results of control procedures in the ALLBUS 1994. In 1994 personal registers from registration offices started being used in the survey as sample frame. The previous sample method was random route (ADM-System, der Heyde and Loeffler (1993)), in which interviewers selected sample units within the two last stages of the selection process.

Table 1: Selected studies dealing with the detection of data fabrication.

Authors	**Survey**	**Share of 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; American population; NHIS	random: 0.2%, combined: 3.6%	X	X		X
Krejsa et al. (1999)	Large scale survey; three American regions; Census 2000 dress rehearsal	random: 0.006%, combined: 0.6%	X	X		
Turner et al. (2002)	Large scale survey; Baltimore population	49% of 451 interviews contributed by 6 falsifiers (net sample 1 200 interviews)	X	X		
Ex-post studies						
Murphy et al.(2004)	Large scale survey; American population; NSDUH	No information		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

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 it is possible 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 3 505 interviews realised in the ALLBUS were controlled by Koch, comparing gender and age of interviewed and selected persons.

- All cases with deviations detected by Koch (n = 196) were controlled by a new contact (in person, by phone or by post). As a result, contact with the interviewer could not be confirmed by sampled persons for 45 interviews. These interviews were declared to be complete falsifications. In 51 other cases someone other than the selected person was interviewed. In yet another 31 cases mistakes were found (these interviews are declared as interviews, in which a part of interview was falsified). Fifty sampled persons could not be reached through controls and for the remainder technical problems leading to deviations were found.

Koch was able to effectively extend the routine control procedures used by the data collection institute and he found a considerable number of cases with deviations. For his analysis he used only information related to the gender and age of target persons. In case of deviation recontact procedures were used. In this way, 2.3% of all interviews realised in the ALLBUS were classified as data falsifications. Koch mentions that the detection method 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. Falsifiers who are more cautious may not be detected by the procedure described by Koch. Thus, the level of 2.3% of detected falsifications represents the lower

boundary for very gross fabrications. Hence, 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)

A study by 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.

A basic idea by Hood and Bushery is 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; Groves et al. (cf. 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 theses 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.

Krejsa et al. (1999) employed a very similar approach. The authors used several indicators based on metadata, such as the proportion of vacant households or dates and times of conducted surveys, in order to define outlier interviewers. These interviewers were then checked during the course of focused reinterviews. Whereas random reinterviews detected only one falsifier out of 1 706 cases, focused reinterviews found 10 falsifiers out of 1 737 cases.

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: 6 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 31% of the time on average. All interviews submitted by these 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) to find suspected 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 authors compared falsified and non-falsified interviews to find clues which could be useful in detecting falsifications. In particular, falsifiers produced implausible data regarding the composition of households (most of the households in falsified interviews included only one 18-45 year old adult) and phone numbers of respondents were not provided for false interviews. Additionally, responses related to sexual behaviour were not plausible in falsified data.

A 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 suspected 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, results 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. Perhaps Koch and other authors (who obtained low or very modest numbers of falsified interviews) would have been more "productive", if they had applied controls to all cases assigned to a dishonest interviewer.

Murphy et al. (2004)

Murphy et al. (2004) analysed data produced by 3 known 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. In the case of the NSDUH, the laptop registered time stamps for each question and each interview step in both modes. This allowed for the calculation of elapsed time for each respective action.

Like Turner et al. (2002) Murphy et al. 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 3 falsifiers were among the top four interviewers, if interviewers were ranked according to their index values. As in the study by Turner et al. (2002) it turned out that falsifiers struggle to adequately reproduce answers to very sensitive questions. Thus, if those questions are available, they might serve as good indicator for constructing a reinterview sample.

Murphy et al. 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 falsifier was generally much faster than the other interviewers, but the other two falsifiers were much slower.

Overall, the study suggests that responses to sensitive questions, rather than lengths of different interview modules, are a more reliable indicator for detecting cheating interviewers. When analysing these responses, it must be kept in mind that emerging patterns may also be due to the allocation of interviewers to certain areas. Considering this, it is important to somehow control for this factor. In terms of the environment in which Hood and Bushery (1997) operated there was no reference study available to the authors, so they stratified their indicators according to some of the respondent's characteristics. Such an approach always bears the risk of omitting some decisive characteristics.

Schräpler and Wagner (2005)

Schräpler and Wagner (2005) examined the data from the German long-term panel study SOEP. In such panel studies fabrications are relatively rare since respondents are interviewed every year and consistency checks across the different waves immediately reveal fraudulent data. By means of two different ex-post analyses Schräpler and Wagner (2005) examined only data from the first waves of different samples of the SOEP (Schupp and Wagner 2002).

The first one was based on the so called Benford's Law (Benford 1938), which we illustrate in more detail in Subsection 3.3. The idea behind this method is to compare the distribution of the first digit of all numbers in the (metric) answers from the survey with the Benford distribution. If the numbers follow a specific monotonic declining distribution, simply spoken that the proportion of 1's is higher than the proportion of the 9's, one can assume that the data is Benford distributed. Schräpler and Wagner (2005) calculated the deviation from the Benford distribution by means of a chi-square value for every interviewer cluster rather than for every single interview. As a result they showed that about half of the known fakers could indeed be marked as cheaters through application of this detection method.

Apart from this analysis Schräpler and Wagner (2005) created an interesting approach by incorporating several variables in order to detect known fakers in survey data. The authors called this the "variability method" because the idea behind it is that cheaters show a lower variance of specific answers across all their conducted interviews than accurate interviewers do. Schräpler and Wagner (2005) attributed the reduction of variance to the proportions of missing answers, of extreme answers in scale questions and of conspicuously consistent answers across specific questions in the questionnaire. Based on the observed variance of interviews Schräpler and Wagner (2005) calculated a plausibility value for every interviewer in order to identify cheaters. If the plausibility was too low an interviewer was considered to be a faker. In this way the authors ranked interviewers with respect to their plausibility values and noticed that almost all of the known cheaters appeared at the top of the ranking. Additionally, they noted that their results were much better than those which are based on Benford's Law. Thus, we can infer from the results that the "variability method" is a more promising way to reveal falsifiers than Benford's Law.

Résumé

Now we would like to summarize the findings of studies presented in this section. For ex-ante studies Table 1 shows that recontact is the most important method for detecting false data. The studies show that random (or unfocused) recontacts enable detection of only an insignificant number of falsified interviews. In all of these studies unfocused recontact procedures were then refined using supplemental information. This information was helpful in identifying additional fabricated interviews. For focused recontacts or reinterviews information about contact outcomes or the number of interviews carried out by one interviewer, referred to as metadata, were used (Table 1). Other simple data analyses (such as those for age and gender of interviewed persons or the proportion of non-minorities interviewed by one interviewer) were helpful in conducting more

focused recontacts or reinterviews. However, choosing one of these methods for ex-ante studies (and their success in detecting falsifications) was highly dependent on the specific circumstances of the study. More concretely, sampling procedures, the survey topic, and the sensitivity of questions are all associated with the usability of a particular detection method. Ex-post data analyses seem to be an effective method of identifying indicators in order to separate false and real data. But more research should be done here to determine the success of such methods, and it should be reiterated that appropriate methods are often bounded to the specifics of a survey (e.g. the sensitivity of questions asked or the proportion of appropriate questions for the application of Benford's Law or the "variability method"). With the help of selected studies presented in this section we aim to show how several methods and combinations of these methods have been used in ex-post and ex-ante studies.

In the next section we carefully examine the different methods and discuss their usability. We also introduce interviewer characteristics as an additional method. Once dishonest interviewers had been detected, different autors (Koch 1995; Turner et al. 2002) then analysed the extent to which interviewer characteristics differ between honest and dishonest interviewers. However, interviewer characteristics were not used as an identification method in these studies. Thus, we disregarded interviewer characteristics in Section 2.

3 Overview of different approaches

As outlined in Section 2 existing literature suggests that the effectiveness of recontact procedures can be increased if they are combined with other indicators. In this chapter we first examine literature related to recontact procedures and subsequently discuss the suitability of other methods that could be applied to create focused reinterview samples.

3.1 Recontact procedures

The most common method of detecting faked survey data is the recontact method. Using this method respondents are recontacted in person, by mail or by telephone after the initial interview in order to verify whether the initial interview actually took place. Below we will focus on this recontact method and the possible problems associated with it.

In spite of the fact that AAPOR (2003) suggests that face-to-face recontact is the most effective method of detecting fraudulent data, the most common recontact method used in surveys involves sending postcards to interviewed persons with an appeal to them to reply. These postcards mainly ask respondents about the time, date, and critical components or topics of the interview, as well as the interviewer's behaviour. A statement about not being interviewed or implausible time and date information may then be considered as indicators of falsifications. In general this method has some questionable factors as shown by Koch (1995) and Hauck (1969). These factors include memory problems and the willingness

of contacted persons to reply, and these are associated with a selectivity bias. Hauck (1969) sent postcards to interviewed and non-interviewed respondents and noticed firstly that only 50% of interviewed respondents returned the postcards. Secondly, he showed that there were race, age and education differences between persons who sent postcards back and those who did not. This implies that cooperative persons who sent the postcards back did not constitute a sample which makes reliable statements about interviewed persons in general. Thirdly, Hauck (1969) found that fourteen out of 100 non-interviewed persons actually stated that they had been interviewed. Thus, it is clear that memory problems or interviewed target persons not sending back postcards decrease the validity of control results obtained by postcards. In particular, with respect to response rates, telephone and/or personal contacts, also referred to as reinterviews, are more effective than contacts using postcards.

Telephone or personal reinterviews were already mentioned in an early work by Case (1971). In market surveys he conducted controls with telephone reinterviews and revealed that about 27% of the interviews in all the studies examined were not conducted properly. Also, the U.S. Bureau of Census regularly checks a randomly selected portion of interviewers (between 2% and 10%) by reinterviewing particular target persons (cf. Bushery et al. 1999). But regarding response and memory problems the reinterview process is limited by sample size and duration (Cantwell et al. 1992). A large number of reinterviews increases costs and a high number of questions posed within a reinterview is a strain for respondents, and consequently biases results. Reinterviews are also limited by the elapsed time following the interview (a delayed survey for further control purposes (reinterview) bears the risk of memory effects). Thus, what is required is a reinterview sample which is large enough to generate significant results but is small enough to keep costs down.

As we have illustrated above (cf. Section 2), ex-ante studies by Hood and Bushery (1997) and Krejsa et al. (1999) (see also Li et al. (2009)) show that a small, and most of all a non-randomly selected, reinterview sample is more effective in detecting cheating interviewers. Thus, we must infer from these empirical findings that "content based reinterviews" perform much better than random reinterviews, and most notably better than sending out postcards. In the following sections we would like to present the prevalent indicators used for the optimal creation of a focused reinterview sample: usage of metadata, Benford's Law and other statistical analyses of survey data, as well as interviewer characteristics.

3.2 Metadata

As already outlined in Section 2, the notion "metadata" comprises different types of information related to the process of data collection, rather than to the collected data itself. Metadata-based indicators used to detect falsifiers can be divided into two groups: indicators based on interviewer's contact outcomes and indicators based on interview processing, such as date and time stamps.

Contact outcomes refer to information related to how many participants re-

fused the interview or how many participants were ineligible for some reason. As outlined above, Turner et al. (2002) were able to detect a large number of fabricated interviews by focusing recontact efforts on interviewers who had shown a suspiciously high success rate, whereas Hood and Bushery (1997) employed the ineligible unit rate as an indicator to put together their focused reinterview sample. In this context it is recommended to control for the characteristics of the area where an interviewer conducts his/her work, as demonstrated by Hood and Bushery (1997). However, due to a lack of reference data this is often not possible.

Date and time stamps can only be recorded if the interview is conducted using a mode which relies on computer assistance. If these are available, they can be used to examine interview length, or the number of interviews completed within one day or within periods in which interviews were conducted. These types of indicators are employed by Bushery et al. (1999), Krejsa et al. (1999) and Murphy et al. (2004).

Given a very limited number of studies it has so far been quite difficult to evaluate how well metadata can be used to detect cheating interviewers. Krejsa et al. (1999) combined both types of metadata-based indicators, which delivered quite promising results. The results of Turner et al. (2002) are quite promising as well, although the results of Murphy et al. (2004) show less promise. In the case of Turner et al. it should also be kept in mind that general success rates were quite low, probably as a result of the high sensitivity of the questions asked in the course of interviews. Consequently, high rates for falsifiers were extremely noticeable.

A large advantage of approaches relying on metadata analysis is that they can be applied to a vast range of surveys. Whenever interviewers are prescribed which persons or households they are to contact, then one can calculate indicators related to contact outcomes. Whenever interviews are conducted with computer assistance, there is the opportunity to record date and time stamps.

3.3 Benford's Law

The largest part of the scarce research regarding the usage of answer patterns to detect fraudulent interviews is related to the usage of Benford's Law for the analysis of metric survey data. With accurate survey data the distribution of the first digit of these metric answers usually follows the so called Benford's Law (Benford 1938), a logarithmic (Newcomb 1881) and scale invariant distribution (Hill 1995). Thus, the probability that the first digit of the numbers is 1 is higher than the probability that it is 9. In general, Benford's Law could be adapted to data without a built-in maximum (Nigrini 1999) and to data which is not composed of assigned numbers like zip codes or bank accounts. For example, Nigrini (1996) and Tödter (2007) have shown that business and financial data in particular follow this monotonic decreasing distribution. Below we present some comprehensive studies which analysed survey data by means of Benford's Law, and we attempt to illustrate whether the authors were successful in identifying fraudulent data.

In order to reveal faked survey data by means of Benford's Law one must ensure that accurate survey data is actually Benford distributed and that faked data is not. If we look at the literature concerning this topic we cannot assume that this is always true. This was shown by Schräpler and Wagner (2003), who made a more in-depth ex-post analysis concerning the raw survey data from the German SOEP. The authors showed that the proportion of the first and second digits of real metric survey data is close to the Benford distribution and that faked data is not. But these findings should be interpreted carefully since the authors also observed a high proportion of 0's and 5's in the accurate survey data, perhaps due to a rounding effect. Therefore, one cannot infer that real survey data is surely Benford distributed. Results from the faked SOEP survey data are in turn not generalisable because of a very low number of cases of fabricated interviews (27 cases for faked data vs. 894 cases for real data). These results are similar to the findings of Wang and Pedlow (2005), who also observed a rounding effect in accurate survey data.

Hence, Schräpler and Wagner as well as Wang and Pedlow, tested some modifications to the analysis in order to improve the detection of cheaters. Schräpler and Wagner calculated a chi-square value for "interviewer clusters", meaning that they pooled together all interviews by one interviewer into one cluster. But the clusters with faked data were not precisely revealed. Again the reasons are a very small number of cases and, most of all, the spatial homogeneity of some clusters. The latter is due to the dependent distribution of interviewers and areas in the SOEP. But even if the authors take this spatial homogeneity into account by means of a linear regression on the chi-square value, faked and real clusters cannot be clearly separated. By contrast – concerning the first digit – Wang and Pedlow calculated the distribution the data set actually follows, the "all cases distribution" (cf. Swanson et al. (2003)). By using this distribution instead of that of Benford's Law, it now became possible for the authors to identify cheating interviewers. Further modifications to the usage of the digits, for example regarding the digit 5 as 0, only led to satisfying results if the authors used the "all cases distribution" as well.

Thus, we infer from the results of Wang and Pedlow and Schräpler and Wagner that real survey data does not clearly follow Benford's Law and that one must be aware of the occurrence of a rounding effect. And there is still a lack of evidence that faked survey data is not Benford distributed, although Diekmann (2007) made an interesting contribution to this topic. Diekmann inspected the first and second digits of unstandardised faked regression coefficients and noticed that they deviate from the Benford distribution concerning only the second digit. But one should notice that this has only been documented for regression coefficients and not for raw survey data. Additionally Diekmann (2010) pointed out that one must avoid stating that survey data which deviate from Benford's Law are automatically falsified. It often occurs that the whole data set follows Benford's Law while some subsamples do not, even if they have not been falsified. The authors supposed that the reason behind this is that the numbers available for the analysis depend on the types of questions and topics in questionnaires. If one disregards this dependent structure, the rate of

"false positives" increases significantly, and therefore the discriminatory power of Benford's Law becomes too low.

To summarize, the current work concerning Benford's Law questions the validity of this method. There is little empirical evidence that real survey data are close to the Benford distribution and problems like rounding effects may lead to a deviation from the Benford distribution. Also, it has not been clearly shown that faked raw survey data are not Benford distributed, since the number of cases is too low in most analyses. Thus, further modifications to Benford's Law, such as several combinations of digits, should be developed and tested in datasets with a higher proportion of falsifications.

Additionally, there is a lack of research regarding the number of interviewers and the number of interviews per interviewer which are required for the identification of falsifications to produce precise results. Recent work by Storfinger and Winker (2011), in which Benford's Law is used as one of four indicators, suggests that the performance somewhat worsens as the number of interviews per interviewer decreases and improves as the overall number of interviewers decreases. However more research is needed to asses how many questionnaires per interviewer are needed in order to successfully adopt Benford's Law. Also, the number and types of variables in the questionnaire which are suitable for application to Benford's Law have not been precisely identified (cf. Porras and English 2004). Is Benford's Law only appropriate for metric variables in raw survey data, or only for statistical estimates like regression coefficients? As Scott and Fasli (2001) demonstrated in a synthetic way, data are more likely to conform to Benford's Law if the dataset contains only positive numbers and is positively skewed with a modal value not equal to zero. But there is still a lack of evidence for real survey data which fit to the Benford distribution. So far Benford's Law can be used to evaluate survey data quality (cf. Judge and Schechter 2007) but it is not efficient enough to precisely identify cheaters.

3.4 Other Statistical Analyses of Survey Data

Inspection of concrete survey data also delivers encouraging results by revealing "at risk" interviewers and falsifications. Several forms of questions and questionnaires could be included in such an analysis.

As a result of our literature review we differentiate between two kinds of indicators which can be applied to an analysis to compare false and real data. The first kind applies to more or less plausible answers to survey questions by falsifiers, while the second kind is associated with the answering behaviour of falsifiers, which may differ from that of real respondents. We will call the first kind of indicators "content related" and the second "formal". Answers to open numerical questions, which were analysed using Benfords Law (cf. Subsection 3.3), also belong to the formal criteria.

Content related indicators are special substantial answer patterns which systematically differ between fraudulent and accurate collected data. In the study by Turner et al. (2002) falsifiers produced implausible data regarding the composition of households (most households in falsified interviews included only

one 18-45 year old adult). Furthermore, in this study target persons "interviewed" by falsifiers "reported" having remarkably more lifetime partners and "showed" substantially more sexual activity, having sex much more often in the past seven days than those who were really interviewed. Finally, falsifiers provided phone numbers of "interviewed" persons less often. Further examples – as found by Reuband (1990) – are differences in evaluating one's own personal financial situation, which was estimated more optimistically by falsifiers than by real sampled persons. However, content related indexes have seldom been used in studies to detect falsified data. Koch (1995) and Hood and Bushery (1997) describe accordant analyses. Hood and Bushery (1997) used the provision of telephone numbers in combination with other indexes in their multiple approach in order to improve reinterview samples. Schräpler and Wagner (2005) compared data from different rounds of a panel study (SOEP) and identified false data in this way. In other studies identified differences between false and real data are only discussed as possible cues to identify falsifiers (Turner et al. 2002). Schnell (1991), who systematically analysed differences between real and falsified data and in particular found differences in correlations and multiple regressions, remains sceptical about the usage of obtained differences for the purpose of detection. Of course, applying content criteria, apart from questions about household composition, age and gender, is problematic since it requires asking sensitive questions (e.g. Turner et al. 2002) or questions about content not related to the survey topic. Application of content related indicators needs additional strong hypotheses and knowledge about the differences between falsified and real data, which are, as a rule, not available.

Formal criteria are produced by analysing the answering behaviour of interviewed persons. An example is item non-response. As shown by Schräpler and Wagner (2005) and Bredl et al. (2008) this occurs less often in falsified than in real interviews, since falsifiers incorrectly assume that real respondents would answer all questions in an interview – and therefore they avoid item non-response. In addition, Bredl et al. (2008) showed that falsifiers differ from real survey respondents in the way they answer semi-open-ended questions when the category "others" was included. Falsifiers tend to avoid the "others" category in order to reduce the effort needed to formulate an open answer. In contrast, when using open-ended questions one has to consider that falsifiers should answer such questions more frequently than real interviewees, but they should tend to shorter answers. In addition, falsifiers are less extreme if they use rating scales for their answers. With the help of multivariate cluster analyses Bredl et al. (2008) were able to separate falsifiers from honest interviewers using (in addition to Benfords Law, see section 3.3) information on item non-response, extreme answers and answering of open and semi-open questions.

Furthermore, cheating interviewers can also be identified by looking at the proportion of answers to questions that lead to a faster answering, and therefore a quicker interview (cf. Matschinger et al. 2005; Hood and Bushery 1997). One would suppose that falsifiers tend to answer these so-called "filter questions" in such a way that allows them to skip a part of the questionnaire "legally" and therefore save time. For a health survey in the US Hood and Bushery

(1997) reported that falsifiers selected "the shortest path through the interview" producing survey participants who live in one person families and are "white non-smoker, no health problems and no health insurance" (p. 821). Using filter questions in this way was one of the multiple indicators applied by Hood and Bushery (1997) to detect falsifiers.

In summary, formal criteria have advantages over content criteria in that they can be used for different types of content in different surveys. Work by Schräpler and Wagner (2005) and Bredl et al. (2008) shows that using formal indicators is an encouraging approach. However, application of formal criteria may increase the complexity of the answering process (e.g. by providing numerous open and filter questions) and reduce the accuracy of respondent's answers by promoting low motivation and superficial information processing (satisficing behaviour, cp. Krosnick and Alwin (1987)). Like content related criteria, falsifiers' answering behaviour should be known to the detectors. Furthermore it is important to ensure that falsifiers cannot adapt their way of cheating easily if the criteria are known to them (e.g. if a falsifier is aware that interviewers are considered to be suspicious if they deliver questionnaires with a low prevalence of item non-responses he/she could simply increase the prevalence of item non-responses when fabricating the data). This problem might be tackled by combining several criteria which makes adaptation from the falsifier's side more difficult. Altogether, more research is needed in order to find an optimal questionnaire form, which can help to detect falsifiers through statistical data analysis using formal criteria.

3.5 Interviewer characteristics

In most of the studies presented above interviewer characteristics were discussed as one issue which can be used for more focused reinterviews. The authors agree that inexperienced interviewers are likely to show cheating behaviour and should consequently be controlled by extended procedures (Biemer and Stokes 1989; Wetzel 2003; Turner et al. 2002; Schreiner et al. 1988). Schreiner et al. (1988), who report the results of the Interviewer Falsification Study of the U.S. Bureau of the Census, discovered by data analysis of two national surveys that the mean duration of employment for falsifiers is significantly lower (1.72 years) than that of all interviewers (6.22 years). The authors recommended that "(...) for the newer interviewers it may be useful to reinterview some of their work more frequently" (Schreiner et al. 1988: p.496). Concerning other interviewer characteristics mixed results can be found in the literature. With respect to gender and age of interviewers Koch (1995) has shown that young interviewers with a higher level of education produce a higher rate of falsifications. In West-Germany Koch found no gender differences, but in East-Germany male interviewers fabricated interviews more often. However Schräpler and Wagner (2003) did not discover any age, gender or education effects in SOEP.

As a result, interviewers' length of service seems to be a sufficient cue for more focused reinterviews. However, experienced interviewers are not less likely to falsify, but are less likely to be detected by controls. Hood and Bushery

(1997) reported the results of a study by the US Census Bureau 1982 which analysed interviewers' characteristics and the likelihood of falsification. The results showed (similar to other studies) that interviewers' tenure in particular is associated with differences in falsification behaviour. But additionally the authors reported that new interviewers with less than five years of experience had a higher probability of being detected, since they falsify more of their assignments and they tend to falsify entire interviews. Experienced interviewers (with five or more years of experience) falsify a smaller proportion of their assignments and prefer to falsify only a part of the interview. As a consequence, falsifications by experienced interviewers are more difficult to detect. Additionally, as shown by Schreiner et al. (1988) experienced interviewers use more selective approaches for falsification than less experienced interviewers. In panel surveys, for example, experienced interviewers falsify more often in continuing households (and not in newly selected households), since data can be appropriately estimated from the previous round.

But in particular the comparison between experienced and inexperienced interviewers is based on very small sample sizes of interviewers who were found to falsify, and thus analyses were often done without statistical tests or results were afflicted by high statisitical insurance.

In summary, newly hired and relatively inexperienced interviewers are seen as requiring extensive controls, but at the same time they are more amateur falsifiers who are more likely to be detected. This topic needs more research, especially regarding methods which can help to identify partially fabricated interviews conducted by experienced interviewers. Using length of service or other demographic characteristics for generating focused interviews appears to be less efficient in light of the results discussed in this section. This is particularly due to the fact that extended controls of less experienced interviewers lead to lower rates of falsification detection for experienced interviewers, who are much more sophisticated falsifiers.

4 Discussion and outlook

In spite of the scarcity of scientific publications related to detection of data fabrication by interviewers, an examination of existing literature has delivered useful insights. Below we attempt to deduce some recommendations for practitioners.

We have distinguished between two types of studies: ex-ante studies describe approaches implemented to detect cheaters, whereas ex-post studies apply indicators to datasets with known cases of falsification. Ex-ante studies analysed in Section 2 clearly suggest that focused recontacts are more effective than recontacts based on random samples. Furthermore, examination of studies using different recontact procedures reveals that the reinterview is the most reliable of these procedures. Forwarding postcards is a questionable alternative as several studies suggest that using postcards leaves many falsifiers undetected. We thus consider the focused reinterview as a good strategy to detect falsifiers.

Ex-post studies complement ex-ante studies in that they provide deeper in-

sights into the suitability of different methods when creating focused reinterview samples. We have defined four methods in this context: metadata, Benford's Law, other analyses of data contained in questionnaires referring to formal or content-related information, and interviewer characteristics. These can be divided into methods available for most surveys and methods whose applicability is only possible for specific types of surveys.

Metadata, most types of formal survey data and data on interviewer characteristics are available for a wide range of surveys. Approaches relying on metadata and formal survey data have delivered promising results when used to detect falsifiers. This is especially true in the case of contact outcome data. Analysis of answer patterns for filter questions seems to be another promising approach and is directly linked to hypotheses of falsifiers' behaviour. Interviewer characteristics did not turn out to be a useful indicator for creating focused reinterview samples. However, literature reveals one important point: experienced interviewers falsify in ways that make their detection more difficult. Thus, indicators relying on metadata or survey data can be expected to deliver better results for inexperienced interviewers who, if they falsify data, do it in a more amateur way. Consequently, it is plausible to assume that focused reinterviews are less efficient when applied to experienced interviewers.

The applicability of Benford's Law and of content related indicators based on substantial answers by survey participants depends on the nature of the data collected. Using Benford's Law requires a wide range of metric variables. Even if a multitude of these variables is available it still remains open whether data which is honestly collected can be supposed to conform to Benford's Law, and whether fabricated data cannot. There have been some promising results, but Benford's Law should be applied with caution and, if possible, in combination with other indicators. General statements about the usefulness of content-related survey data are difficult to make, as the type of data available depends on the type of survey. It can be stated that answer patterns to questions on very sensitive issues have turned out to be good predictors of interviewer cheating and should thus be used to detect cheaters when available.

Whenever one employs indicators based on metadata or content-related survey data one has to keep in mind that striking indicator values are not necessarily caused by data fabrication but may also be the result of "conventional" interviewer effects or cluster related design effects (spatial homogeneity). If indicators for the creation of focused reinterviews are calculated on the interviewer level it is not possible to distinguish between data fabrication and interviewer effects. Several studies which attempt to control for spatial homogeneity have been discussed, however these approaches are not always replicable and they bear the risk of omitting decisive factors. It is possible that indicators based on formal survey data – for example the proportion of extreme answers or item non-response – are less impaired by conventional interviewer effects and spatial homogeneity. However, to the best of our knowledge no research has yet been done on this issue.

In addition to the recommendations we have provided above, we would like to point out some ideas for further research. First of all, more research should be

done to broaden the discussed findings with respect to the suitability of methods we presented in Section 3. Of course there are several studies which focus on one or more of these methods, but as we have shown in the context of Benford's Law the results are sometimes inconclusive or are based on a sample which is too small. Thus, further studies are needed to gain more reliable insights into how every single detection method performs, as well as how combinations of these methods perform. This means that we need to learn whether a combination of detection methods performs better than the application of only one method.

In particular, the usage of survey data, and especially several content-related and formal indicators, poses the problem of a trade-off between the complexity of questionnaires and their usability on the part of respondents. Following our recommendation that one should apply different detection methods, the questionnaire to be examined should provide a high proportion of suitable questions to derive several indicators for the statistical analyses of survey data. In doing so, the complexity of the questionnaire may increase and therefore the usability for the interviewer, and of course for the respondent, diminishes and produces further biases. Thus, one must keep in mind that the questionnaire overcharges neither interviewers nor respondents.

Concerning the structure of survey data examined and regardless of the detection method used, we noticed that in the majority of the studies presented, analyses were conducted only on the interviewer level. Yet identification of fraudulent survey data could also be realized on the interview level (cases in a data set). Focusing on interviews bears the problem that the amount of data per interview is less than the amount of data per interviewer, which probably reduces discriminatory power. On the other hand, approaches focusing on interviewers might struggle to detect falsifiers who fabricate only a small proportion of their interviews. Which strategy is preferable is another topic for future research. Finally, it should be kept in mind that the methods discussed above have all been tested and developed primarily to identify fabrications of entire interviews. Further work concentrating mainly on the identification of partially faked interviews is also needed.

The literature overview presented here shows that first conclusions can be drawn from existing comprehensive studies about interview falsifications and the prevailing detection methods. However, most of the approaches – especially in the analysis of collected data – should be evaluated through further research related to their effectiveness, and approaches should be developed to enable an exact detection of falsifications and falsifiers.

References

AAPOR (2003). Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects.
Available under: http://www.aapor.org/Content/NavigationMenu/ResourcesforResearchers/falsification.pdf(Access: 18.11.2010).

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society 78*(1), 551–572.

Bennett, A. (1948). Toward a solution of the "cheater problem" among part-time research investigators. *Journal of Marketing 12*(4), 470–474.

Biemer, P. and S. Stokes (1989). The optimal design quality control sample to detect interviewer cheating. *Journal of Official Statistics 5*(1), 23–29.

Bredl, S., P. Winker, and K. Kötschau (2008). A statistical approach to detect cheating interviewers. ZEU Discussion Paper Nr. 39.

Bushery, J., J. Reichert, K. Albright, and J. Rossiter (1999). Using date and time stamps to detect interviewer falsification. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 316–320.

Cantwell, P., J. Bushery, and P. Biemer (1992). Toward a quality improvement system for field interviewing: Putting content reinterview into perspective. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 74–83.

Case, P. (1971). How to catch interviewer errors. *Journal of Advertising Research 11*(2), 39–43.

Crespi, L. (1945). The cheater problem in polling. *Public Opinion Quarterly 9*(4), 431–445.

der Heyde, C. V. and U. Loeffler (1993). Die ADM-Stichprobe. *Planung und Analyse 5*, 49–53.

Diekmann, A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics 34*(3), 321–329.

Diekmann, A. (2010). Benford's law and fraud detection. facts and legends. ETH Zurich Sociology Working Paper No. 8.

Durant, H. (1946). The "cheater" problem. *Public Opinion Quarterly* (2), 288–291.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology*. Wiley. New Jersey.

Harrison, D. and S. Krauss (2002). Interviewer cheating: Implications for research on entrepreneurship in Africa. *Journal of Developmental Entrepreneurship 7*(3), 319–330.

Harrisson, P. (1947). A british view on "cheating". *Public Opinion Quarterly 11*(1), 172–173.

Hauck, M. (1969). Is survey postcard verification effective? *Public Opinion Quarterly 33*(1), 117–120.

Hill, T. (1995). A statistical derivation of the significant digit law. *Statistical Science 10*(4), 354–363.

Hippler, H. (1979). Untersuchung zur Qualität absichtlich gefälschter Interviews. ZUMA Arbeitspapier.

Hood, C. and M. Bushery (1997). Getting more bang from the reinterviewer buck: Identifying 'at risk' interviewers. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 820–824.

Judge, G. and L. Schechter (2007). Detecting problems in survey data using benford's law. *Journal of Human Resources 44*(1), 1–24.

Kennickell, A. (2002). Interviewers and data quality: Evidence from the 2001 survey of consumer finances. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 1807–1812.

Koch, A. (1995). Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994. *ZUMA-Nachrichten 36*, 89–105.

Köhne-Finster, S. and G. Güllner (2009). Ergebnisse der Interviewerbefragung im Mikrozensus. *Wirtschaft und Statistik 5*, 397–405.

Krejsa, E., M. Davis, and J. Hill (1999). Evaluation of the quality assurance falsification interview used in the census 2000 dress rehearsal. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 635–640.

Krosnick, J. and D. Alwin (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly 51*(2), 201–219.

Li, J., J. Brick, B. Tran, and P. Singer (2009). Using statistical models for sample design of a reinterview program. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 4681–4695.

Matschinger, H., S. Bernert, and M. Angermeyer (2005). An analysis of interviewer effects on screening questions in a computer assisted personal mental health survey. *Journal of Official Statistics 21*(4), 657–674.

Murphy, J., R. Baxter, J. Eyerman, D. Cunningham, and J. Kennet (2004). A system for detecting interviewer falsification. Paper Presented at the American Association for Public Opinion Research 59th Annual Conference.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics 4*(1/4), 39–40.

Nigrini, M. (1996). A taxpayers compliance application of Benford's Law. *Journal of the American Taxation Association 18*, 72–91.

Nigrini, M. (1999). I've got your number. *Journal of Accountancy 187*(5), 79–83.

Porras, J. and N. English (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 4223–4228.

Reuband, K. (1990). Interviews, die keine sind. "Erfolge" und "Mißerfolge" beim Fälschen von Interviews. *Kölner Zeitschrift für Soziologie und Sozialpsychologie 42*(4), 706–733.

Schnell, R. (1991). Der Einfluss gefälschter Interviews auf Survey Ergebnisse. *Zeitschrift für Soziologie 20*(1), 25–35.

Schräpler, J. (2010). Benford's Law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP). SOEP Papers.

Schräpler, J. and G. Wagner (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 Series, 969.

Schräpler, J. and G. Wagner (2005). Characteristics and impact of faked interviews in surveys – an analysis of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv 89*, 7–20.

Schreiner, I., K. Pennie, and J. Newbrough (1988). Interviewer falsification in census bureau surveys. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 491–496.

Schupp, J. and G. Wagner (2002). Maintenance of and innovation in long-term panel studies: The case of the German Socio-Economic Panel (GSOEP). *Allgemeines Statistisches Archiv 86*(2), 163–175.

Scott, P. and M. Fasli (2001). Benford's Law: An empirical investigation and a novel explanation. CSM technical report, Department of Computer Science, University Essex.

Stewart, N. and S. Flowerman (1951). An investigation of two different methods for evaluation of interviewer job performance. *Personnel Psychology 4*(2), 161–170.

Storfinger, N. and P. Winker (2011). Robustness of clustering methods for identification of potential falsifications in survey data. ZEU Discussion Paper Nr. 57.

Sudman, S. (1966). New approaches to control of interviewing costs. *Journal of Marketing Research 3*(1), 56–61.

Swanson, D., M. Cho, and J. Eltinge (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 4172–4177.

Tödter, K.-H. (2007). Das Benford Gesetz und die Anfangsziffern von Aktienkursen. *WiSt 36*(2), 93–97.

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

Wang, Y. and S. Pedlow (2005). Detecting falsified cases in scf 2004 using Benford's Law. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 3652–3657.

Wetzel, A. (2003). Assessing the effect of different instrument modes on reinterview results from the Consumer Expenditure Quarterly Interview Survey. In *Proceedings of the American Statistical Association (Survey Research Methods Section)*, pp. 4508–4513.