Statistical Journal of the IAOS 32 (2016) 295–303
DOI 10.3233/SJI-161012
IOS Press

Assuring the quality of survey data: Incentives, detection and documentation of deviant behavior

Peter Winker
Department of Statistics and Econometrics, University of Giessen, Licher Strasse 64, 35394 Giessen, Germany
Tel.: +49 641 9922640; E-mail: Peter.Winker@wirtschaft.uni-giessen.de

Abstract. Research data are fragile and subject to classical measurement error as well as to the risk of manipulation. This also applies to survey data which might be affected by deviant behavior at different stages of the data collection process. Assuring data quality requires focusing on the incentives to which all actors in the process are exposed. Relevant actors and some specific incentives are presented. The role of data based methods for detection of deviant behavior is highlighted as well as limitations when actors are aware of them. Conclusions are drawn on how settings can be improved to provide positive incentives. Furthermore, it is stressed that a proper documentation of data quality issues in survey data is required both in order to increase trust in the data eventually used for analysis and to provide input for the development of new methods for detection of deviant behavior.

Keywords: Data fabrication, incentives, deviant behavior

1. Introduction

Empirical research in the social sciences and research data are interwoven almost tautologically. Therefore, the quality of research outcomes depends substantially on the quality of research data be they primary data, i.e. collected for the research project, or of secondary nature. It is common knowledge that not all data are what they pretend to be or what researchers expect them to be. This might be due to classical measurement errors, not appropriate instruments, sample selection etc., but also due to deviant behavior of people involved in the collection and compilation of data. Some recent examples are manipulations by the researchers leading to retraction of an article in Science [26], doubts about official statistics regarding national debt [13,28], and interviewers or supervisors duplicating and fabricating data [4,6,23] to name just a few examples. According to anecdotal evidence, these and other reported cases might rather represent the tip of the iceberg (see also the meta-analysis by Fanelli [14]).

This contribution focuses on deviant behavior in the data collection process and its effect on data quality in surveys. As long as conditions for conducting a survey are not perfect, there is a risk of all types of deviations from prescribed routines by all agents involved in the process, e.g., substituting respondents, using shortcuts in the questionnaires, (partial) fabrication by the interviewer, or duplication of all or part of the questionnaire by the supervisor. Obviously, it is important to use all available means to reduce such deviations to an absolute minimum. It might be less evident why one should also report on the effort spent on data quality and possibly conspicuous findings. However, in order to maintain confidence in empirical research both a detailed description of measures ensuring data quality and reports about conspicuous cases is essential. The latter reports are also relevant input for the development and improvement of methods aiming at identifying deviant behavior. It has to be accepted that there is not a safety limit up to which some tampering of data is not harmful. In fact, e.g., Schraepler and Wagner [30] have shown that even the inclusion of a very small number of fabricated data might have a big im-

1874-7655/16/$35.00 © 2016 – IOS Press and the authors. All rights reserved
This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0).

pact, in particular for multivariate analysis. And the effects might reach far beyond scientific analysis, as survey results might drive policy decisions and, consequently, fabrication could end up in a misallocation of public funds [21].

Three central aspects regarding data quality in surveys will be discussed in this paper. First, in Section 2, a closer look will be given to the incentives provided to all agents involved in the collection of survey data. In particular, it is discussed to what extent these incentives are supportive for high data quality or not. Second, in Section 3, the role of data based methods for the identification of potential fabrication is scrutinized under the predator and prey perspective, i.e., whether they will improve data quality if agents are aware of their usage. Third, the necessity of providing proper documentation of all types of deviations detected is indicated, which is linked to the issue of reputation building in Section 4. The final Section 5 provides some concluding remarks.

2. Incentives

Figure 1 shows some actors involved in the production of survey data. It shows that the collection of survey data requires the interaction of several parties starting with the researcher defining the research questions, selecting instruments and a sampling scheme down to the respondent answering some questions in an interview and back through the data cleaning and aggregation procedures. Obviously, data quality is affected by decisions at all levels. The situation is complicated by the fact that in most cases, the specific actions are not observed. This implies that principal (e.g. a supervisor) and agent (e.g. an interviewer) do not share the same information about the actual activities of the latter. In a context of asymmetric information, setting incentives right becomes even more relevant in order to ensure high data quality. We will discuss some of the decisions at the individual level, which incentives they are subject to, and to what extent asymmetric information is relevant. However, we cannot claim to provide the full picture at all levels of this process.

Starting with the researcher, some cost and benefit of specific actions can be described. First, developing a questionnaire, which is easy to handle both by the interviewers and the respondents, and at the same time useful for collecting the data required for a particular analysis, takes time. Second, having the survey done involves cost in form of the payment to the field agency. Additional effort might be spent on controls of field work or checks of data quality ex post. The benefits come, e.g., in form of gaining scientific insights, publishing research papers or, more general, an improved reputation as a researcher. Unfortunately, cost and benefit are not linked closely. For example, only a small number of journals require making research data publicly available [12]. Consequently, careful checking of survey data and – possibly – repeating (parts) of the survey in case of doubtful findings comes at a cost for the researcher (time and payment to field agency) without contributing directly to publication results and overall reputation. Only her own scientific interest actually sets the incentives right towards achieving high data quality. In many or most cases, this is sufficient – in some, it is not.

When selecting a field agency, the researcher operates as a principal. She would prefer to receive high quality data in exchange for the payment, which, however, should be as low as possible. Obviously, the researcher cannot always observe in detail how the field agency is operating, e.g. how they select and train interviewers,[1] which methods are used to supervise interviews and to control incoming data. Such a situation of systematic asymmetry in information might result in a market break down. The researcher may not be willing to pay enough for research firms to conduct high quality surveys as she cannot be sure to obtain high quality given the risk of undetected deviating behavior. As a consequence, agencies operating at high cost with good quality might have to leave the market. Anticipating this negative selection will make the researcher even more reluctant to pay a high enough price – the classical situation of a "market for lemons" results [1]. The only mechanism supporting a more favorable outcome is reputation built up over time by the field agency allowing it to charge higher prices for higher quality. However, a part from anecdotal evidence, no formal process of reputation building, e.g. by ratings as on some online trading platforms, is established so far.

For the field agency, cost and benefit are also multidimensional. The agency incurs cost in form of effort for supervision and controls as well as in form of the payment for interviewers, while benefit comes both through payment by the researcher and reputation which might have a positive effect on future contracts

[1] A noteworthy exception is the PIAAC study described in Massing et al. [25], when the researchers have been involved directly in the training and supervision of interviewers.

Fig. 1. Involved parties in data collection process.

and prices. A part from the potential impact on reputation, incentives are not set in a way to foster high data quality. As long as data quality issues do not become public, it appears rational for the field agency not to discuss them with the researcher nor to invest effort on repeating (parts) of the survey at its own cost. It is also rational to pay interviewers per completed interview although it is known that payment per hour reduces the risk of deviant behavior and improves the quality of collected data as does a careful training and supervision of interviewers [16,25,33].

There is less common knowledge about the role of supervisors in the process of collecting data. However, it seems obvious that they also have to spend effort on training and supervision of interviewers as well as on screening incoming data. Furthermore, they might be under pressure to reach certain quota. For them, it might be most difficult to build up a reputation. At the same time, given that a substantial number of classical control procedures such as re-interviews or phone calls are organized through the supervisors, it might be less risky for a supervisor fabricating data than for an interviewer.

The choice of interviewers is also subject to asymmetric information similar to the selection of field agency by the researcher. Typically, no public knowledge is available regarding interviewers' past performance, in particular with respect to potential deviations from prescribed protocols. If such misconduct were observed, most likely the interviewer would have to leave the agency, but might find easily a new position if accepting the contract offered. Furthermore, apart from active supervision, it is difficult to find out what exactly interviewers are doing in the field and how carefully they follow the protocol. Therefore, the incentive structure imposed on them deserves a detailed analysis.

The discussion on setting the incentive structure for interviewers dates back at least to Crespi [8] who discussed "demoralizers" of different type having a negative effect on the quality of interviews. Such demoralizers include fixed quota making it more likely to replace a respondent by someone else, payment per completed interview, sensitive questions, long or repetitive questionnaires. Further cost involve travel time to sampling points, risky environments [19] or reduced payment if quotas are not met. Unless the interviewer has some intrinsic interest in the study at hand, his main benefit is the payment received.[2] There is hardly potential to build up a positive reputation which might increase employment prospects or payment structure within a field agency and even less so for a transferable positive reputation. As a consequence, it has to be expected that payment per completed interview increases the number of interviews done during a specific span of time be it by being less meticulous in following all detailed instructions, by deviating behavior, e.g. by exploiting options in the filter scheme, or by generating partial or complete falsifications. Similar outcomes might result if interviewers are not paid for travel to distant sample points or if risky environments are not taken into account in an appropriate way. Controls and data based detection methods might counteract these effects to some extent, in particular if they are hard to "predict" for an interviewer. However, as long as the main risk the interviewer faces if he is caught is that he might have to change the employer, a substantial detection probability is required to assert data quality.

To sum up, we are faced with the "market for lemons" situation at several stages of the data collection procedure. In such a situation, for different actors in the process it might be rational behavior to betray trust. As a consequence data quality might be reduced, which is not in the interest of the researcher as long as her scientific interest is concerned. However, even for the researcher, reputation and options for publications may not be actually impacted as long as data quality is not challenged! And given that it does not matter, the researcher will go for the cheapest offer as does the field agency with regard to the interviewer!

Summarizing the discussion about incentives and asymmetric information at the interviewer level, it seems that enabling the building up of reputation might be a key element to improve the situation if positive reputation pays out in form of higher payment and better working conditions for the interviewer. Currently, reputation can be built by an interviewer only while working for a specific institute.

[2] For a discussion of cost and benefit for the interviewer see also Kennickell [21].

To improve the situation, one idea would be to implement a central register for interviewers collecting information about the number of surveys and interviews they conducted including information about confirmed or suspected deviant behavior. Data security is a critical issue for such a register. In principle, even the researcher should have access to this information on "her" interviewers, but at the same time, anonymity not only of the interviewer but possibly also of respondents (due to a correlation of region and interviewer) might be affected. Some more thought and research is needed to come up with a sensible proposal on how to set up such a register, to set incentives for all involved parties to contribute to the register and to use it. After all, similar instruments have been developed successfully for other applications such as peer to peer lending [15,34].

3. Detection of deviant behavior

Considering the description in the previous section, it is quite evident that the incentives provided currently in the process of survey data collection are not optimal with respect to ensuring high data quality. The problem of asymmetric information amplifies the issue. It has been pointed out above that a mechanism for reputation building has to be established to evade this unpleasant setting. However, building up reputation also requires methods for detecting deviant behavior. Otherwise, even having conducted a large number of interviews (for an interviewer) or surveys (for a field agency) would not increase reputation. We will come back to the issue of reputation building and the role of documentation of suspicious cases in the following section. Here, we concentrate on methods for detection of deviant behavior mainly at the interviewer level, although some of the approaches might also be applied at other stages of the process.

How could we identify deviant behavior, in particular partial or complete fabrications at the interviewer level? Which type of tool will be most valuable in terms of increasing incentives for following prescribed procedures assuming that some kind of reputation building is feasible? Before turning to some methods proposed in this context, two general remarks appear appropriate. First, the cost of application of these methods have to be taken into account. Obviously, a one-to-one follow up by the scientist herself could eliminate all suspicious cases, but such an approach would neither be feasible given the high cost involved nor incentive compatible in practice. Therefore, the cost and – potential – benefit of different methods have to be put in perspective. Second, we have to take into account that people involved in the survey process might learn about the methods in place and adjust their behavior accordingly. This is the predator and prey perspective. Two examples might illustrate the issue. For the first example, consider more or less straightforward duplicates of real interviews, e.g. added by the interviewer or supervisor to complete specific quota. If it was known that duplicate checking procedures[3] are in place, diligent fabricators would add some random noise to their data to reduce the risk of detection. The second example refers to the so called "Benford's law" [2] describing a property of the distribution of first digits of metric data often observed, which is used, e.g., both in the context of survey data [6,20] and in accounting [11,29] to detect fraudulent behavior. Again, if interviewers are aware that data will be checked by comparing the distribution of first digits with the theoretical distribution according to Benford's law, interviewer could adjust to this benchmark when fabricating data with some effort. However, Benford's law is not a law of nature, and metric data as they come up in surveys may or may not conform with it. Consequently, choosing as benchmark the actual empirical distribution of first digits for real data would be both a more sensible criterion and, given that this true distribution is unknown for all agents including the interviewer until the survey is completed, a much more challenging one for a potential fabricator. While shortly discussing some methods used for detection of potential falsifications in survey data, the two aspects will be addressed.

Before turning to the detection methods, it should be retained, that in most cases a final decision on the quality of a specific interview can be made only after a re-interview with the respondent.[4] Given the cost involved in this procedure, it appears sensible to focus on interviews conducted by "interviewers at risk", i.e. interviewers more likely of having deviated from the protocol [18]. The selection of these "interviewers at risk" might be done focusing on interviewers' characteristics or on properties of the collected data, meta- or para-data.

[3] For example, the method proposed by Kuriakose and Robbins [23].

[4] It was found that alternative methods, e.g. contact by phone or postcards generate a high rate of errors of type one and two, i.e. respondents might remember a visit by an interviewer, which has never taken place or was quite short just for collecting basic socio-demographic information, or respondents might have forgotten about a visit which actually has taken place. See, e.g. Hauck [17] and Cantwell et al. [7].

In their review, Bredl et al. [5, p. 18f] point out that the only interviewer characteristic found systematically linked to the risk of deviating behavior is experience (for an example see Lawrence and Love [24]). More experienced interviewers are found more seldom to deviate from prescribed routines or even to fabricate data. This might be a result of reduced cost of conducting interviews according to the protocol due to experience or an effect of informal reputation gained with a specific field agency possibly resulting in a wage premium. However, it could also be an artefact due to the fact that more experienced interviewers know control procedures better and, consequently, avoid being detected more easily. Therefore, focusing on specific interviews based on interviewer characteristics does not appear a very promising approach. Nevertheless, it should be kept in mind that some – typically unobserved – characteristics might affect the quality both of real and fabricated data.[5]

During the last few years, several proposals have been made aiming at identifying "interviewers at risk" based on collected data. It is not intended to provide a review of all these approaches, but rather to focus on some classes of methods, namely content, duplicates, uni- and multivariate distributions of (meta-)indicators and para-data.

In content based methods, values of core variables of the survey (e.g. gender or age) collected by one interviewer in his interviews are compared either with their actual values (in case of register based samples) or with the population frequencies. While the method is rather easy to implement, the extent to which suspicious interviews might be found appears limited. In fact, it has been reported repeatedly that fabricated data match quite well univariate distributions of classical socio-demographic variables [30]. Furthermore, if interviewers conduct only a small number of interviews, the comparison with population frequencies might result in a large number of false alarms. Finally, if interviewers are aware that checks are in place based on such criteria, it should not be too difficult for experienced interviewers to avoid detection.

Also methods based on identifying duplicates or near-duplicates in interviews collected by one interviewer or provided by a single supervisor might not be a robust method under the predator and prey perspective. As pointed out above, if it is known that duplicate checking is in place, adding some random noise to the duplicates might be sufficient to avoid detection. At the same time, the instruments would have to be tuned more aggressively to still find at least some of potentially suspicious cases, which, again, would result in a large number of false alarms.[6]

Moving from the data themselves to some meta-indicators might render the task more difficult for falsifiers. Such meta-indicators have been used already for some time. Typical examples are the already mentioned distribution of first digits for metric data, the frequency of item-nonresponse, or the frequency of choosing the category "other" in semi-open questions. The difficulty for a potential fabricator consists in the fact that the "true" value of these indicators is not known a priori and might depend substantially on the specific survey both due to the construction of the questionnaire and the choice of sample. Consequently, even if interviewers are aware of controls based on these indicators, they will find it difficult to fabricate data with the required properties. However, this advantage also turns out to be a disadvantage for the practical implementation of the method. First, a substantial number of data from real interviews has to be collected to obtain an estimate of the value of such meta-indicators for the specific situation, and only afterwards, selection of "interviewers at risk" may start. The method proposed by Bredl et al. [6] tries to tackle this situation by using a clustering procedure to split up the interviewers in two groups, the supposedly honest ones and those "at risk". Obviously, this approach is also subject to the risk of errors of type one and two. However, it appears more robust to the predator and prey perspective as fabricators would have to know the "true" values of the indicators. In fact, in their approach several indicators are used simultaneously in a multivariate setting, which renders the task for the potential fabricators even more difficult as they would have to match the unknown multivariate distribution of these meta-indicators while generating data.

Finally, if interviews are done computer-assisted, para-data, e.g. on duration of interviews, key typing speed, waiting times between entering answers (for reading the questions) or even GPS-data could be exploited. Clearly, the process of fabricating data would become more challenging in such a setting. Nevertheless, if interviewers are aware of the methods, they could still replace respondents or just fill out the questionnaire with a friend or colleague at the right co-

[5] Winker et al. [33] provide evidence on such effects in an experimental setting.

[6] For a discussion of false alarms resulting from duplicate checking see also Simmons et al. [31].

ordinates. At least, the cost of generating fabricated data would increase, possibly making it more attractive to do the real interview instead. Unfortunately, these methods, in particular the GPS-data approach, might not be feasible in not register-based surveys and for surveys under risky conditions, when the risk of falsifications appears to be highest.

To sum up this short survey of different approaches targeted at detecting fraudulent interviewers, it is found that some of the methods proposed might provide sensible signals only if potential fabricators are not aware of their use, while others can be used only under specific circumstances. Relying on multivariate meta-indicators appears to be most promising under the predator and prey perspective. However, finding the best meta-indicators for this purpose is not an easy task and also depends on the specific survey. Therefore, all methods should be analyzed in more detail on a larger set of test cases, which might include but should not be restricted to simulated or experimental data only.

4. Documentation and reputation

As pointed out above, a more comprehensive documentation about identified cases of deviant behavior in the survey process is essential for at least three purposes. First, while estimates about the prevalence of deviating behavior are mainly based on anecdotal evidence up to now, such a documentation would result in a sensible estimate of a lower bound of the frequency of deviant behavior. Second, datasets comprising identified falsifications provide the testbed for the development of new and the evaluation of existing method for detecting deviant behavior. Third, any mechanism of reputation building in the context of asymmetric information requires some feedback mechanism which requires information about detected cases.

Addressing the first purpose appears straightforward, but requires a good coverage of all types of surveys and either incentives for all involved persons to report every case of deviant behavior or a project for randomly selecting a representative set of surveys to undergo an in depth analysis. It might be expected that the prevalence of deviant behavior depends on characteristics of the survey, e.g., whether it is register based or not, the mode of the survey, the length of the questionnaire, whether sensitive questions are asked or not etc. However, so far, only anecdotal evidence is available on these topics.

The second purpose is even more obvious. As mentioned by Koczela [22, p. 283]: "Without better sharing of datasets that include fabricated interviews, progress in fabrication detection methods will be hindered." Although there exists some sporadic evidence when and how interviewers or supervisors might deviate, these findings do not point towards uniform behavior. It rather seems that the specific way of deviant behavior and its impact on collected data depends on interviewers and the setting of interviews. Thus, for a full picture of deviant behavior, which might be used to improve and evaluate detection methods, access to many more real datasets with identified deviant behavior is required. When it comes to evaluate the performance of specific detection methods for specific types of deviant behavior, experimental and simulated data sets might be used to complement the findings from real data [3,31,33].

Although they cannot replace the access to real datasets with identified fabrications, experimental datasets also exhibit some advantages. First, the design of the survey can be adjusted in all facets to the specific question at hand. For example, to study the impact of asking sensitive questions, in an experimental setup, the same questionnaire with and without these questions can be used in an otherwise identical setting. Based on such standardized conditions, it becomes possible to generate even more synthetic data by means of bootstrap simulation procedures with well defined properties [9,10,32]. However, these advantages come with some limitations. First, also experimental surveys are costly. Given that the generated data can be used solely as a testbed for detection methods and not for substantial analysis, funding might be an issue. Second, an experimental setting remains an artificial setting. Due to financing constraints, often only students are recruited as interviewers and to guarantee a controlled setting also the respondents might be students recruited for the experiment [27].

The third suggested use, i.e., input for a reputation building mechanism, is probably the most controversial. Again, based solely on anecdotal evidence, it seems that reporting on identified deviant behavior in a survey risks to affect reputation of the field agency, the research outcome or even the researcher negatively. Therefore, even when deviant behavior was found and the compromised data have been cleared, e.g., by repeating parts of the survey or simply be removing the suspicious data, this is not reported to the public, i.e., cannot be found in final publications. Again, incentives in place do not operate in the right direction. While researchers with some experience in this domain would see it rather as a positive signal for data quality if

some details are provided on detection methods used and how suspicious cases have been dealt with, for the broader (scientific) public any hint towards problems in data collection contributes to negative reputation.

When thinking on how to reduce asymmetric information and provide means to build reputation, the interaction between researcher and field agency might not be the easiest part of the process to deal with for the reasons just mentioned. However, eventually, it should become common practice and established standard to also report about the methods used for identifying deviant behavior and suspicious or confirmed cases of deviant behavior found in a dataset. Reputation building could start with joining a voluntary agreement between field agencies to provide full information including datasets with identified fabrications at least to interested researchers in a joint data research center guaranteeing data protection both with regard to the survey content and the involved parties. Such a network might be established by research organizations and incentives for field agencies might be set by restricting access of public research funding to projects involving participating firms.

For the agent at the final stage of the process, i.e., the interviewer, a sensible mechanism for reputation building requires some incentives for participation. In particular, it should come along with some financial benefits in form of a wage premium for certified interviewers which would have to be transferable when working with another agency. The central register for interviewers discussed in Section 2 above might provide such a framework. Of course, it would have to be reliable, i.e., only interviews which have been collected under a documented data quality strategy should be considered, and a mechanism is needed ensuring that identified deviant behavior is marked in the register. Finally, a wage premium for registered interviewers might be obtained by a requirement for public funded surveys to make use solely of registered interviewers.

Obviously, these ideas of establishing reputation building mechanisms for improving data quality are just a first attempt. A more detailed analysis of such institutional settings is required checking also the incentives for participating and disclosing information in such a framework.

5. Conclusions

High data quality in empirical social research cannot be taken for granted. The incentive structure as it is in place now exhibits substantial shortcomings increasing the risk of deviant behavior. Effort is needed both in research and survey practice. More research is needed regarding details of the incentive structure, its impact on data quality and improved mechanisms of data collection and quality management. Survey practice should aim at improving the benefit to cost structure for interviewers and other key actors in the data collection process. Existing knowledge about interviewer behavior already suggests that imposing fix quota and using a payment per completed interview scheme increase the risk of deviant behavior, while careful interviewer training, long term contracts and payment per hour are reducing such a risk.

Typically, information at different stages of the survey process will not by symmetric. Consequently, there is a potential for deviant behavior to reduce cost. Thus, methods for detecting suspicious cases have to be developed and applied. Different approaches focusing on meta- and para-data as well as on data based methods should be used as complements to identify suspicious cases. Follow-up interviews are required to eventually decide on the status of an interview found to be suspicious by such methods. The methods have to be robust in the sense that they should be able to provide relevant signals also if agents, e.g. interviewers, are aware of their usage. This calls for approaches focusing on properties of interviews which are not known a priori to the interviewer and, possibly, also not to the supervisor and field agency. In particular, considering a set of such properties and their multivariate distribution will make it very tricky for an interviewer to fabricate interviews matching this unknown multivariate goals.

For the development and evaluation of methods for detecting deviant behavior, more real datasets with identified falsifications and other deviations from the protocol have to be made available for research. However, this is only one reason why a more detailed documentation of such cases in datasets, which are used for research or policy advice, is required. If no information about control procedures, suspicious cases found and action taken on them is provided, one might conclude that no such procedures have been in place, which might create doubt about the quality of the dataset and the results obtained from it. In contrast, a proper documentation allows the user of data and research results to assess to what extent data quality is assured. Thus, it is not information about critical cases which should create doubt, but rather the absence of any information on quality control procedures and their results.

Finally, implementing detection methods and reporting on the results is a necessary component for the es-

tablishment of a framework for signalling and reputation building of interviewers, supervisors, field agencies and – indirectly by selecting "good" partners – also for researchers. Beyond this necessary input, additional components are required which might comprise, e.g., participation in cost and time intensive certified training courses for interviewers, the option to register for an "interviewer-id" which will be linked to each dataset generated by the interviewer combined with the obligation for surveys run with public funding to employ trained and registered interviewers only. Developing and putting in place such a framework will increase the trust in survey based research which is crucial for empirical social sciences and many other fields.

Acknowledgements

The paper is based on a presentation given at the NORC Fabrication/Curbstoning Conference at Bethesda, MD, on February 23rd 2016. I am indebted to the organizers, Fritz Scheuren and Steve Koczela for having given me the opportunity to present and discuss some results of my research on the identification of falsifications in survey data, to Steve Koczela for his valuable feedback on a preliminary version of this paper, as well as to participants of the workshop for their valuable comments and suggestions. Financial support through the DFG in project WI 2024/5-4 is gratefully acknowledged.

References

[1] G.A. Akerlof, The market for "lemons": Quality uncertainty and the market mechanism, *The Quarterly Journal of Economics* **84** (1970), 488–500.

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

[3] B. Birnbaum, G. Borriello and A.D. Flaxman, Brian DeRenzi and A.R. Karlin, Using behavioral data to identify interviewer fabrication in surveys, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13. ACM. New York, NY, USA, 2013, pp. 2911–2920.

[4] J. Blasius and V. Thiessen, Detecting poorly conducted interviews. In: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, P. Winker, N. Menold and R. Porst, eds, 2013, pp. 67–88. Peter Lang. Frankfurt.

[5] S. Bredl, N. Storfinger and N. Menold, A literature review of methods to detect fabricated survey data. In: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, P. Winker, N. Menold and R. Porst, eds, 2013, pp. 3–24. Peter Lang. Frankfurt a. M.

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

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

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

[9] S. De Haas and P. Winker, Identification of partial falsifications in survey data, *Statistical Journal of the IAOS* **30** (2014), 271–281.

[10] S. De Haas and P. Winker, Detecting fraudulent interviewers by improved clustering methods – the case of falsifications of answers to parts of a questionnaire, *Journal of Official Statistics* p. forthcoming, 2016.

[11] C. Durtschi, W. Hillison and C. Pacini, The effective use of Benford's law to assist in detecting fraud in accounting data, *Journal of Forensic Accounting* **5** (2004), 17–34.

[12] M. Duvendack, R.W. Palmer-Jones and W.R. Reed, Replications in economics: A progress report, *Econ Journal Watch* **12** (2015), 164–191.

[13] European Commission (2010). Report on Greek government deficit and debt statistics. Technical report. European Commission. Brussels.

[14] D. Fanelli, How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data, *PLoS ONE* **4** (2009), e5738.

[15] S.M. Freedman and G. Zhe Jin, Learning by doing with asymmetric information: Evidence from prosper.com. Working Paper 16855. National Bureau of Economic Research, 2011.

[16] P.A. Gwartney, Mischief versus mistakes: Motivating interviewers to not deviate, in: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, P. Winker, N. Menold and R. Porst, eds, 2013, pp. 195–215. Peter Lang. Frankfurt.

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

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

[19] S. Hornbeck, D. Peng, C. Studner and D. Jodice, Ensuring data quality in conflict zones. Technical report. D3 Systems, 2010.

[20] G. Judge and L. Schechter, Detecting problems in survey data using Benford's law, *Journal of Human Resources* **44**(1) (2009), 1–24.

[21] A. Kennickell, Curbstoning and culture, *Statistical Journal of the IAOS* **31** (2015), 237–240.

[22] S. Koczela, Discussion of Identification of partial falsification in survey data, *Statistical Journal of the IAOS* **30** (2014), 283–284.

[23] N. Kuriakose and M. Robbins, Don't get duped: Fraud through duplication in public opinion surveys, *Statistical Journal of the IAOS* p. forthcoming, (2016).

[24] C. Lawrence and E. Love, Characteristics of falsified interviews. In: *Proceedings of the American Statistical Association (Survey Research Methods Section)*, 2010, pp. 4771–4780.

[25] N. Massing, D. Ackermann, S. Martin, A. Zabal and B. Rammstedt, Controling interviewers' work in PIAAC – the programme for the international assessment of adult competencies, in: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, P. Winker, N. Menold and R. Porst, eds, 2013, pp. 117–130. Peter Lang. Frankfurt.

[26] M. McNutt, Editorial retraction. *Science*, 2015.

[27] N. Menold, P. Winker, N. Storfinger and C.J. Kemper, A method for ex-post identification of falsifications in survey data, in: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, P. Winker, N. Menold and R. Porst, eds, 2013, pp. 25–47. Peter Lang. Frankfurt.

[28] T. Michalski and G. Stoltz, Do countries falsify economic data strategically? *Some Evidence That They Might* **95** (2013), 591–616.

[29] M. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley. Hoboken, NJ, 2012.

[30] J.-P. Schräpler and G.G. Wagner, Characteristics and impact of faked interviews in surveys – an analysis of genuine fakes in the raw data of SOEP, *Allgemeines Statistisches Archive* **89**(1) (2005), 7–20.

[31] K. Simmons, A. Mercer, S. Schwarzer and C. Kennedy, Evaluating a new proposal for detecting data falsification in surveys. Technical report. PewResearchCenter. Washington, DC. last accessed: 4/28/2016, 2016.

[32] N. Storfinger and P. Winker, Assessing the performance of clustering methods in falsification using bootstrap, in: *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, P. Winker, N. Menold and R. Porst, eds, 2013, pp. 49–65. Peter Lang. Frankfurt a. M.

[33] P. Winker, K.-W. Kruse, N. Menold and U. Landrock, Interviewer effects in real and falsified interviews – results from a large scale experiment, *Statistical Journal of the IAOS* **31** (2015), 423–434.

[34] J. Yan, W. Yu and J.L. Zhao, How signaling and search costs affect information asymmetry in p2p lending: The economics of big data. *Financial Innovation* **1**(1) (2015), 1–11.