A Method for ex-post Identification of Falsifications in Survey Data

Peter Winker[1], Natalja Menold[2], Nina Storfinger[3], Christoph J. Kemper[4], Sabrina Stukowski[5]

[1] Justus-Liebig-University Giessen, Center of International Development and Environmental Research (ZEU), Giessen, Germany,
e-mail: Peter.Winker@wirtschaft.uni-giessen.de

[2] Gesis – Leibniz Institute for the Social Sciences, Survey Design and Methodology, Mannheim, Germany, e-mail: Natalja.Menold@gesis.org

[3]Justus-Liebig-University Giessen, Center of International Development and Environmental Research (ZEU), Giessen, Germany,
e-mail: Nina.Storfinger@zeu.uni-giessen.de

[4] IMPP, Mainz, Germany, e-mail: contact@christoph-kemper.net

[5] Gesis – Leibniz Institute for the Social Sciences, Survey Design and Methodology, Mannheim, Germany, e-mail: Sabrina.Stukowsi@gesis.org

Abstract

Interviewers have a substantial impact on data quality. Their motivation to deviate from prescribed routines is analyzed. Thereby, the focus is on falsifications of survey data. Based on approaches from cognitive psychology and principal agent theory data based indicators are constructed which should differ between real and falsified interview data. A multivariate cluster analysis is applied to a set of such indicators to identify interviewers who are more likely to have contributed falsified data and might be subject to a follow up in a fieldwork setting. A heuristic optimization algorithm is used for the clustering instead of sequential procedures. Data obtained from an experiment are used to evaluate the performance of the indicators and of the multivariate method. The experiment used two payment schemes for the interviewers – per interview and per hour. It is also analyzed to what extent the payment scheme affects interviewers' behavior with regard to falsifications.

Keywords: Interviewer behavior; Clustering; Data quality

Acknowledgements: Financial support from the DFG through projects WI 2024/3-1 and ME 3538/2-1 within SPP 1292 is gratefully acknowledged. We are indebted to Michael Blohm, Sebastian Bredl, Gesine Güllner, Marie Kristin Opper, Rolf Porst and Viktoria Trofimow for their helpful support.

1. Introduction

Survey data quality might be negatively affected in many different ways. Several issues, such as sampling, non response, wording and structure of questionnaires or motivation of interviewees have been subject to intensive analysis. In this contribution, we focus on another source of contamination of survey data, namely deviant behavior of the interviewers in a face to face setting, in particular the situation when interviewers fill in part of or even complete questionnaires themselves instead of properly conducting the full interview. Although there exists substantial anecdotal evidence on the prevalence of this problem, statements about data problems related to interviewers' deviant behavior are seldom found in the literature. A possible reason for this lack of coverage might be the associated risk of generating doubt about the data quality. Also contributions dealing with method for identification of deviant behavior are rather scarce (see Bredl et al. (2013) for a literature review).

On the other hand, it appears obvious that even a small proportion of falsifications in survey data might affect further empirical analysis to a non trivial extent. Given the interviewers' knowledge about the surveyed population, it is not too surprising that estimates of unconditional means or variances are often found to be only marginally affected by falsifications. However, as shown, e.g. by Schräpler and Wagner (2003), even a small proportion of fabricated data can be sufficient to cause strong biases in multivariate statistics.

The issue of interviewers' deviant behavior is taken into account in actual survey practice. However, reliable methods for detecting falsifications are expensive, e.g. the random re-interview. Thus, it is important to apply such methods to a subsample of the data comprising those data which are most likely to have been subject to some type of falsification. We present an approach to generate such focused re-interview samples making use only of the available data from the survey. Of course, if additional information is available, e.g. metadata, it might be used to complement the analysis.

In this paper, we present recent results of a German Research Foundation (DFG) funded research project dealing with ex-post detection of falsified data in face-to-face surveys. While the methodological approach builds on the multivariate indicator based cluster method suggested by Bredl et al. (2012), a closer theoretical analysis of interviewers' motivation is used to identify indicators which might help to discriminate between falsifiers and interviewers actually conducting their interviews. Our hypotheses concerning these indicators are mainly based on theories of cognitive psychology and of respondents' (assumed) motivation to answer survey questions. Furthermore, we assume that falsifiers use stereotypes about potential respondents, since detailed individual data about respondents' opinion and behavior are not available to them. Finally, also aspects from principal agent theory are taken into account. As a result the set of potentially useful indicators becomes much larger than the four indicators discussed by Bredl et al. (2012).

Besides extending the set of indicators and improvements of the clustering method, we also report on the results of three experimental studies run in the framework of the research project (Menold et al. 2013). While the first two studies, which might be considered as a pretesting phase, are only briefly sketched, the main focus of this paper is on the third experiment with a large number of interviewers and conducted interviews. This experiment also included a variation of the payment scheme for the interviewers. A part from reporting the findings from these experiments, we will also provide some comments on the robustness of the results.

The remainder of this paper is structured as follows. Section 2 presents the theoretical background used to derive indicators of interviewers' deviant behavior and describes those indicators. Section 3 describes the experiments and the results obtained for the indicators. The multivariate clustering procedure is introduced in Section 4, which also provides results of its application to the experimental data. Major conclusions and an outlook to further research are presented in Section 5.

2. Theoretical Framework and Indicators

The proposed method relies on finding data based indicators which help to discriminate real data from those produced by falsifiers. For the selection of such indicators it is important to understand interviewers' motivation with regard to deviant behavior and how they actually act when producing falsifications. Given the lack of reports on falsifications in the literature, we have to resort to behavioral theories which might be useful for describing deviant behavior in this context. We complemented the first two pretesting experiments with cognitive interviews to find out more about the actual motivation and procedure of falsifying interviewers. The theoretical analysis and the selection of indicators also builds on the insights gained from these cognitive interviews.

While there is a substantial amount of literature on respondents' behavior in an interview setting, the behavior of interviewers appears to be less well studied. Nevertheless, some results regarding respondents' behavior, e.g., motivated by results from cognitive psychology (e.g., Tourangeau et al. 2000) or by research on respondents' motivation to answer survey questions (Krosnick and Alwin 1987) can be adapted for the interviewers' perspective as well. In fact, we assume that interviewers are confronted with a decision problem where their actual action cannot be easily observed, i.e., a typical principal agent situation. Thus, they face a tradeoff between not following all instructions given and the – possibly – high cost in doing so. The cost of cooperation include working time (including travel time to respondents' household), demoralisers such as bad or too long questionnaires, asking sensitive questions or visiting risky neighborhoods. When deviating, they are faced with a (perceived) risk of detection and the associated cost (dismissal, criminal proceedings). Obviously, the interviewer can influence the risk of detection by the way he or she falsifies data. It is our aim to understand how this is done and whether statistical methods can help to uncover such falsifications.

We continue with the assumption that if interviewers decide to deviate, they aim at reducing the risk of being discovered by providing faked data which satisfy the survey agency. Consequently, the "satisficing model" introduced by Krosnick and Alwin (1987) with regard to respondents can also be used as frame to explain the differences between real and falsified data which then serve as a source for indicators to identify falsified data. In this setting, satisficing describes a behavior trying to minimize cognitive effort when filling in the questionnaire. As a consequence, we might expect more often the selection of the "do not know" category if available or response patterns such as extreme or middle responding in item batteries. Furthermore, we assume that falsifiers might resort to use stereotypes (Hippler 1979; Schnell 1991) about potential respondents, e.g., when detailed individual data about respondents' opinion and behavior are not available to them.

When considering data based indicators which might proof helpful for the classification of interviewers, we distinguish two types of indicators, formal and content-based. Formal indicators are based on differences in response behavior between real respondents and falsifiers. This class has the major advantage that it depends just on the questionnaire structure, in particular, on the type of questions asked, i.e. allows for a straightforward application in different surveys. In principle, it should be even possible to develop tools for constructing questionnaires which at the same time also generate procedures for the calculation of such formal indicators once the data are collected. Content-based indicators focus on the differences regarding the substance of responses to specific survey questions. Therefore, content-based indicators depend heavily on the particular survey setting. Consequently, they appear to be less versatile in their application. Nevertheless, some general ideas might be easily adopted across a wide variety of surveys in social and economic sciences.

Let us start with a short description of some formal indicators which are used in our empirical application. Given the assumptions about the motivation of falsifiers, we might expect the avoidance of the "others" category in semi-open questions (Bredl et al. 2012) as well as responding to filter questions in a way to avoid subsequent questions (Hood and Bushery 1997). Consequently, in the case of filter and semi-open questions we expect falsifiers showing higher satisficing (less effort) than real respondents. This is measured by the indicators SEMI-OPEN (relative frequencies of choosing the "others, please specify" category) and FILTER (frequency of choosing the option allowing to skip part of the questionnaire), respectively. However, falsifiers might also reduce satisficing (higher effort) in regard to other types of questions. These opposed tendencies result from the two conflicting motivations of falsifiers: to save time and to avoid detection. In combining these two conflicting tendencies falsifiers try to save time, only if a legitimate response by reduced effort is possible. Accordingly we expect less satisficing in regard to response tendencies, i.e. extreme and midpoint responding style, acquiescence, rounding, primacy and recency effect. Furthermore, we expect less item non-response (INR) overall (Bredl et al. 2012; Shaeffer et al. 2005) and with regard to open-ended questions (OPEN). Another indicator is given by the variances. We expect lower variances in the falsified data. This is due to less extreme responding by falsifiers and the usage of stereotypes to compensate for a lack of information about the respondent (Reuband 1990, Schnell 1991).

Content-related indicators used in the study are the frequency of choosing non-existing response options in the Vocabulary and Overclaiming Test (VOCT, Ziegler et al. 2013) and in a question about magazines read on a regular basis. For both questions part of the given response options were fictive. Another content-related indicator for falsification found in our previous studies is, that falsifiers tend to underestimate past political activity of respondents.

Due to space constraints, we cannot discuss and present all indicators in detail here. They can be found in Menold et al. 2013.

3. Experimental Evidence

The development and evaluation of the indicators was based on two preliminary experimental studies (Menold, Storfinger and Winker 2011; Menold and Kemper 2011). For these studies the experimental design was similar to the one proposed by Hippler (1979). Confirmed data from the German General Social Survey (ALLBUS) 2008 are used as benchmark. Mainly students were asked to generate falsified data. They received a small set of socio-demographic indicators, behavioral and opinion information from real ALLBUS and were asked to complete the questionnaire themselves. To develop our indicators and to test our assumptions concerning different indicators, we subsequently compared the real ALLBUS data and the data produced by the "falsifiers". Furthermore, cognitive interviews have been conducted with some falsifiers to obtain a better understanding of strategies used for generating the falsified data.

The results presented in this paper are based on the third experimental study based on a much larger sample. Both real and falsified data were collected within the experimental setting. Thus, a specific questionnaire could be used including the type of questions necessary to construct the indicators which have been found useful in the preliminary steps of the analysis. For the experiment, N=78 students have been recruited. In a first phase, they completed each about 10 real face-to-face interviews. In order to make sure that all interviews were actually conducted, they have been tape recorded and all recordings have been controlled. In a second step, the interviewers had to produce fabricated survey data in the lab. Thereby, the procedure was the same as the one described above, i.e. they obtained some basic information about one of the real respondents (of course, from a person interviewed by someone else) and had to complete the questionnaire. In this way, we obtained a data set of N = 710 falsified interviews corresponding to each of the N = 710 real interviews.

The experiment is set up in a way to mimic a scenario with experienced interviewers as the students run 10 real interviews prior to start working on the falsifications. Furthermore, there was an incentive given to produce "good" falsifications. If their falsifications were not detected by our multivariate method, students could gain a price. Finally, we used two different treatments with regard to remuneration of the interviewers. In one group, payment was per completed interview, while it was per hour in the second group.

To test whether indicators are sensitive to falsification, we conducted a between-subject Multivariate Analysis of Variance (MANOVA) of false vs. real interviews. In addition, we compared the differences between real and falsified data concerning indicators used on the aggregated interviewer level as they are used for the cluster analysis. As a result, we found significant differences for all except one indicator between real and false data. In most cases, the sign of the difference points in the direction expected based on theoretical reasoning.

We found significant less extreme responding, less acquiescence, as well as lower primacy and recency effects in falsified data than in real data. The results for INR showed fewer unanswered items on the part of the falsifiers, although this result is not significant. Next, falsifiers used filter questions to avoid responding additional questions significantly more often and provided responses to "other, please specify" category in semi-open-ended questions significantly more seldom than real respondents. In these instances, they reduced their cognitive effort by using legitimate options provided in the questionnaire. In contrast to our expectations, open-ended questions were more seldom completed by falsifiers. They also exhibit more often a middle responding style in rating scales.

Also some content-based indicators differ significantly between real and falsified data. In contrast to real respondents, falsifiers identified more often not existing terms in the vocabulary test (VOCT) and used more often fictive newspaper titles in providing information about reading behavior. Finally, falsifiers strongly underestimated real participants' past political participation.

The different payment schemes applied in the experiment are found to affect the outcome for some indicators. For example, it is found that INR was lower if the participants were paid per hour and not per completed interview. At the same time, the difference between real respondents and falsifiers for this indicator becomes more pronounced for the payment per interview condition although it still did not become significant. Also the number of reported past political activities increased in both groups when payment per hour was provided. And while extreme responding does not differ in a significant way between real respondents and falsifiers in the payment per hour setting, the difference becomes highly significant in the payment per completed interview setting.

When concentrating on the effects of the payment scheme for falsifiers only, we find that falsifiers tend to answer the questions in a way that they could extent the interview duration (and therefore increase their total paying) when paid per hour. For instance, falsifiers significantly use more often filter questions to skip a part of the questionnaire when paid per hour than when paid per completed interview.

Overall, the results point out that motivation of interviewers is a relevant factor which can explain the differences between real and falsified data. In addition, the identified effects of payment methods should be taken into account both for the application of multivariate methods to identify falsified data and for further theoretical analysis of interviewers' behavior.

4. Cluster Analysis

Although most of the indicators exhibit significant differences between real and false data, a reliable identification of falsifications based on individual indicators does not appear feasible given a substantial overlap between both groups. Thus, either a high share of false positives (i.e. interviewers erroneously assigned to the group of falsifiers) or a high share of undetected falsifiers will result (see Menold et al. 2013, p. 41). Consequently, the multivariate method proposed by Bredl et al. (2012) is applied. The idea of this cluster procedure is a grouping of the interviewers into two groups, where one group should comprise the honest interviewers and the other the supposed falsifiers. Obviously, also with such a multivariate approach a perfect grouping is not to be expected in real applications. Therefore, we assess the performance of the clustering procedure by investigating the number of correctly assigned interviewers and the number of false positives and false negatives.

While the results in Bredl et al. (2012) are based on a traditional sequential optimization algorithm, we follow their suggestion to use a method aiming at a globally optimum cluster. This could be achieved by Bredl et al. (2012) by means of full enumeration of all possible cluster solutions given the small number of interviewers involved. However, for our experimental setting, the globally optimum cluster has to be approximated by means of a heuristic optimization approach. We make use of an implementation of the threshold accepting (TA) algorithm for this purpose (Winker, 2011). Instead of using a sequential fusion method as in hierarchical clustering, it relies on an iterative improvement of a given solution with regard to a given fitness criterion. The criterion used in the present analysis is the sum of squared distances within each cluster. Thus, the algorithm starts with a randomly assignment of the interviewers, represented by the vector of indicator values corresponding to their interviews, into two groups. Then, in each iterative step, one element is randomly selected and transferred to the other cluster. If this modification improves the fitness of the solution or at least does not reduce fitness by more than a predefined threshold, it is accepted. It can be shown that this algorithm can provide the globally optimum cluster for a given fitness criterion for appropriate parameter settings (threshold sequence) and a number of search steps going to infinity.

Using all available indicators, two clusters result. One comprises 70 interviewers including 61 falsifiers, i.e. 78% of all 78 falsifiers in the sample, while the other cluster has 86 elements including 69 honest interviewers, i.e. 88% of all honest interviewers. Given that the method does not make use of any a priori information about the interviewers, not even about the number of supposed falsifications, the result is quite convincing. Focusing follow up interviews on the “at risk” cluster would definitely improve the survey quality substantially. We also considered subsets of indicators, in particular those exhibiting the highest discriminatory power in the univariate analysis. However, no further improvements, i.e. with regard to both false positive and false negatives could be obtained except of a small improvement when just excluding INR. Thus, it appears that, in general, exploiting all information available in the indicators by means of a multivariate analysis is a promising route for a data based identification of potential falsifications in survey data.

We also applied the cluster method to two subsamples of the data. The first one comprising all interviews conducted under payment by completed interview, the second under payment by hour. The results are qualitatively similar to the full sample for the first case, when 75% of all falsifiers and 85% of all real respondents are correctly assigned. In contrast, the payment per hour setting allows for an even more clear-cut separation of the two groups. In this scenario, 84% of the falsifiers and 95% of the real respondents are correctly assigned. This improved performance might be due to the fact that honest interviewers and falsifiers differ more noticeably in some indicator values when paid per hour. For instance, as mentioned above, falsifiers and honest interviewers produce a lower share of non-response when paid per hour. But falsifiers reduce their share more than honest interviewers resulting in an increasing difference.

It might come as a surprise that detection of falsifications becomes easier when both honest interviewers and falsifiers receive incentives to spend enough time on the questionnaires (payment per hour). However, it has to be noted that in the experimental setting the decision to falsify was imposed exogenously. In a real setting, a payment by hour scheme might not only influence the quality of falsifications, but also the decision on whether or not to deliver real data. This aspect has to be taken into account both for future experiments in this field as well as for a further analysis of falsifiers' motivation and behavior.

There are a number of straightforward extensions of the analysis, which is subject of our current research. First, one might analyse the performance of the clustering method for different objective functions or when a fixed limit on the size of the falsifier cluster is imposed. Both ideas can be easily implemented in the heuristic optimization setting. The goals of imposing a sequentially increasing number of elements in the falsifier cluster would be to assess whether the share of false positives is smaller at the beginning of the procedure and whether a point of saturation might indicate the actual number of falsifications in the sample even more precisely than the unrestricted cluster approach. Second, it will be analyzed how the performance deteriorates when interviewers decide to falsify only part of the questionnaire.

5. Conclusion

Interviewers might deviate from prescribed routines when cost of adhering becomes too high. The motivation to deviate is reflected in statistical properties of the generated data which can be used for identifying "at risk" interviewers. The experimental evidence has shown a high discriminatory power of some of the proposed indicators and an even better performance of a multivariate clustering approach making simultaneous use of several indicators.

Considering a large number of indicators in a multivariate setting has the additional advantage that it becomes quite difficult if not impossible for the deviant interviewer to reproduce the multivariate distribution of these indicators in his falsified data even if he is aware of the indicators used. In particular, if indicators are based on a comparison of the values for one interviewer with those of the remaining ones (and not with a fixed benchmark as, e.g., in the case of Benford's law), a replication will become impossible. Consequently, the method should stay effective even if the interviewers are aware of the technical details of the construction of indicators and the multivariate clustering method.

Obviously, further research is needed to find out to what extent the results of this study can be generalized. For example, in real settings the share of falsifiers is (hopefully) much smaller than in our experiment. A simulation based analysis by Storfinger and Winker (2013) suggests that the discriminatory power of the method improves for a smaller share of falsifiers in the data set. Nevertheless, further experiments and – if accessible – analysis of real data is necessary to confirm this tendency.

Finally, it is strongly recommended to put more emphasis on the motivation and behavior of interviewers both in research and field work. Their contribution is essential for obtaining a satisfactory data quality. The incentive structure for the whole process of collecting data has to be studied and – where necessary – modified in order to improve data quality and integrity for empirical research in the social sciences.

References

Bredl, S., Storfinger N. and Natalja M. (2013). A Literature Review of Methods to Detect Fabricated Survey Data. In Survey Standardization and Interviewers' Deviations – Impact, Reasons, Detection and Prevention, (eds.) P. Winker, P., N. Menold, and R. Porst, Frankfurt a.M, 3 – 24.

Bredl, S., Winker, P. and Kötschau, K. (2012). A Statistical Approach to Detect Interviewer Falsification of Survey Data, Survey Methodology, 38, 1, 1-10.

Hippler, H.-J. (1979). Untersuchung zur "Qualität" absichtlich gefälschter Interviews, ZUMA Discussion Paper, Mannheim, 1979.

Hood, C.C., Bushery, J.M. (1997). Getting More Bang from the Reinterview Buck, Identifying "at risk" Interviewers, Proceedings of the American Statistical Association, Survey Research Methods Section 27, 820-824.

Krosnick, J.A., Alwin, D.F. (1987). An Evaluation of a Cognitive Theory of Response-order Effects in Survey Measurement, Public Opinion Quarterly, 51, 201-219.

Menold, N., Kemper, C.J. (2011). Survey Response Characteristics as Indicators for Detection of Falsifications, Paper presented at the 4th Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland, 2011, July.

Menold, N., Storfinger, N., Winker, P. (2011). Development of a Method for Ex-post Identification of Falsifications in Survey Data, Proceedings of New Techniques and Technologies for Statistics - NTTS 2011, Brussels.

Menold, N., Winker, P., Storfinger, N. and Kemper, C.J. (2013). Development of a Method for ex-post Identification of Falsifications in Survey Data. In Survey Standardization and Interviewers' Deviations – Impact, Reasons, Detection and Prevention, (eds.) P. Winker, P., N. Menold, and R. Porst, Frankfurt a.M.: Peter Lang, 25 – 47.

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

Schäfer, C., Schräpler, J.-P., Müller, K.-R., Wagner, G.G. (2005). Automatic Identification of Faked and Fraudulent Interviews in the German SOEP, Journal of Applied Social Science (Schmollers Jahrbuch), 125(1), 183-193.

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

Schräpler, J.-P. and Wagner, G.G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys. An Analysis by Means of Genuine Fakes in the Raw Data of SOEP, IZA Discussion Paper Series, 969.

Storfinger, N., Winker, P. (2013). Assessing the Performance of Clustering Methods in Falsification using Bootstrap. In Survey Standardization and Interviewers' Deviations – Impact, Reasons, Detection and Prevention, (eds.) P. Winker, P., N. Menold, and R. Porst, Frankfurt a.M.: Peter Lang, 49-65.

Tourangeau, R., Rips, L.J., Rasinski, K.A. (2000). The Psychology of Survey Response, Cambridge: Cambridge University Press.

Winker, P. (2011). Optimization Heuristics in Econometrics, Applications of Threshold Accepting, Chichester: Wiley.

Ziegler, M., Kemper, C.J., Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T), Journal of Individual Differences, in press.