A Method for Ex-Post Identification of Falsifications in Survey Data

Natalja Menold, Peter Winker, Nina *Storfinger, Christoph* J. Kemper

Abstract

Falsified survey data can have a substantial impact on data quality. A method is presented here which enables identification of falsifications in survey data. This method is based on a detailed analysis of the motivation behind falsifications, which permits us to derive hypotheses about the properties of data generated by falsifiers. Using these hypotheses indicators of specific statistical properties of falsified interviews are constructed. Using false and real data generated in an experimental setting, we examine which indicators are valuable in terms of their capability to distinguish between falsifiers and non-falsifiers. Classification of interviewers is based on a multivariate analysis. The results are discussed in relation to methodological issues which arise when this approach is applied in real surveys.

Keywords: Interviewer falsifications; Indicators for falsifications; Classification

Acknowledgements

We would like to thank the DFG for its financial support in projects WI 2024/3-1 and ME 3538/2-1 within SPP 1292. We would also like to thank Michael Blohm, Sebastian Bredl, Gesine Güllner, Marie Kristin Opper, Rolf Porst and Viktoria Trofimow for their helpful support.

Identification of Falsifications in Survey Data

Interviewers may be faced with incentives to deviate from prescribed routines for several reasons. These include difficulties in making contact with the target person, complex questionnaires, or a low payment level based only on the number of completed interviews. Moreover, deviations can take place in a number of different ways. Rather subtile forms include surveying a member of a household other than the intended one, or conducting the survey by telephone when face-to-face interviews are required (Groves et al. 2004). Complete or partial fabrication of interviews can be considered to be the most severe form of deviation or falsification.[1] Whilst research on falsifications of interviews is rare in literature,[2]

[1] The act of fabricating entire interviews is called "curbstoning" by the US Bureau of the Census.

[2] It is more common to remove suspicious data prior to further analysis without reporting in detail on the methods used for identifying interviewers at risk or the share of data removed. Notable exceptions are the NKPS, for which the codebook (http://www.nkps.nl/CodeBook/CodeBookFrame.htm) reports in section 11.2 on the procedure and detected

anecdotal evidence indicates a substantial prevalence of the problem, with estimates typically ranging between 1 and 9 %. However, in certain contexts falsification rates may be even higher, for example as reported by Bredl et al. (2012) for a small survey in a non-OECD country.

Fabricated interviews can have serious consequences for statistical analyses based on survey data. Schnell (1991) and Schräpler and Wagner (2003) provide evidence that the effect on univariate statistics may not be so severe. But even a small proportion of fabricated interviews can be sufficient to cause significant biases in multivariate statistics. Schräpler and Wagner (2003) analysed data from the German Socio Economic Panel (SOEP) and found that the inclusion of fabricated data in a multivariate regression reduces the estimated effect of years of education on log gross wages by approximately 80%, even though the share of fabricated interviews was less than 2.5%. Consequently, methods aimed at preventing and/or detecting falsified interviews are of crucial importance in face-to-face survey research.

The most effective way to identify falsified interviews is the re-interview (Biemer and Stokes 1989). However, for reasons of expense it is impossible to re-interview all respondents in large surveys. Therefore, we must ask how the re-interview sample can be optimized to best detect dishonest interviewers. Generally speaking it seems practical to re-interview respondents who have already been interviewed by an interviewer who appears likely to be a falsifier. Hood and Bushery (1997) use the term "at risk" interviewer in such a context. The problem of determining which interviewers are "at risk" was already addressed in the 1980s, although literature on this issue is still scarce. Previous research was aimed at interviewer characteristics signalling "at risk" interviewers. For example, interviewers with a relatively short length of service have been found to be more likely to falsify interviews (Biemer and Stokes 1989; Koch 1995; Schreiner et al. 1988; Wetzel 2003).

Another approach is to check the plausibility of the data collected by interviewers. Hood and Bushery (1997) used several indicators to identify "at risk" interviewers in the National Health Interview Survey (NHIS). For example, they calculated the rate of households labelled ineligible or the rate of households without telephone numbers per interviewer and compared the rates to census data. Koch (1995) used deviancies in data relating to gender and age of interviewed persons, which were known to the surveyor from official personal registers.

cases, and the German General Social Survey (ALLBUS), for which identification methods and the number of suspicious cases are reported as well.

Additionally, systematic differences between real and falsified data can be used as criteria to identify “at risk” interviewers. We will refer to such criteria as “indicators” here. Thus, falsifiers may produce distributions of opinions and behaviour which differ from those of real respondents. We refer to such indicators as “content-related”. Reuband (1990) found that falsifiers produce more optimistic predictions for future individual economic situations than real respondents. In addition, differences in response behaviour can also be used as indicators. Bredl et al. (2012) found that falsifiers tend to avoid the category “others, please specify” when answering semi-open-ended questions. Schäfer et al. (2005) showed that falsifiers produced less item non-response and less variability in their data. We refer to such indicators, which are not based on differences in substantial responses to survey questions but rather on differences in response behaviour, as “formal” indicators.

The multivariate method proposed by Bredl et al. (2012) simultaneously uses information related to several formal indicators to detect “at risk” interviewers. This method uses only information contained in the data and no other external sources. In doing so, the information from different formal indicators and their dependency structure are used to classify interviewers, for example by means of cluster analyses. This is expected to result in higher selectivity in terms of identifying interviewers “at risk”.

The goal of the current study is to build on the method suggested by Bredl et al. (2012) to distinguish between real and falsified data. We aim to do this by introducing additional indicators and improving methods for multivariate data analysis. This research will provide new methods for identifying data falsifications and is intended to complement other methods aimed at quality control of interviewers’ work (e.g. re-interviews). Our analysis is based on experimental data in which the correct assignment of real and falsified interviews is known. This allows for evaluation of the outcome of the procedure. The dataset is also substantially larger than that studied by Bredl et al. (2012), resulting in greater power of statistical data analysis.

In the next section we will describe the different indicators. The differences between real and falsified data in terms of these indicators are based on the psychology of survey response and approaches to the motivation of falsifiers. Subsequently, we will provide a description of our procedure. Next, we will present the results concerning the ability of indicators to differentiate between falsified and real data. Finally, we will present conclusions based on our results.

Indicators Sensitive to Falsifications

To identify falsifications we do not rely on individual characteristics of interviewers, but rather on indicators derived solely from the survey data. These indicators should help to distinguish between falsifiers and non-falsifiers. In this section we present the theoretical background which helps to provide an explanation for differences between real and falsified data that have been used as sources of indicators in previous research (e.g. Bredl et al. 2012; Hood and Bushery 1997; Shaeffer et al. 2005). Furthermore, these theoretical approaches are helpful in identifying additional content-related or formal indicators which have not yet been used to detect falsified data.

In terms of content-related indicators we consider differences in information availability between real respondents and falsifiers. In particular, when producing substantial responses to survey questions falsifiers tend to make use of stereotypes related to potential respondents (Hippler 1979; Schnell 1991) since little detailed individual information about the respondents is available to them. Stereotypes are theories about characteristics of a social group shared by another social group (Stroebe et al. 1992). When using stereotypes individuals overestimate differences between social groups as well as similarities within social groups (Tajfel 1969). The next idea in terms of content-related indicators is to deliberately exploit the different levels of information availability between real respondents and falsifiers. Here a special kind of control question comes into play. For example, one can ask "What magazines do you read?" and show a card presenting a list of magazines. Considering that half of the magazines on the list do not actually exist, falsified data would deviate from real data (Noelle-Neumann 2003). In this case it is less likely that real respondents would chose fictitious magazines because exact information is available to them and they only have to look for the magazines in the list they actually. Falsifiers may show a higher probability of choosing a non-existent magazine since they do not expect that some response alternatives are implausible. We assume that indicators based on other fictitious response options produce similar differences between real and falsified data.

When considering formal indicators we rely on the "satisficing model" introduced by Krosnick and Alwin (1987). The authors define satisficing as a superficial kind of information processing in which respondents minimize their cognitive effort when responding to survey questions. Here respondents do not try to find an optimal answer, but rather just an acceptable one. For example, satisficing results in frequently selecting the "do not know" category, in often providing extreme or neutral positions, or – in extreme cases – in providing the same response to every item (straight lining). We assume that real respondents and falsifiers differ in their satisficing behaviour, thus producing differences in formal indicators.

In most cases the predominant motivation of falsifiers is to save time and effort, which consequently results in a higher level of satisficing when compared to real respondents. This can explain falsifiers' avoidance of the "others, please specify" category in semi-open-ended questions, which has been found by previous research (Bredl et al. 2012). The "others, please specify" category was used in questions with a list of nominal response options. A response which does not match with any other available option should fit into the "others, please specify" category. Since the "other" information is known to the real respondent and is easy to provide, this person may have a low tendency to exhibit satisficing behaviour in this setting. In contrast, in the case of falsifiers providing additional information when choosing the category "others, please specify" is associated with a higher level of cognitive effort, since falsifiers then have to create this additional information. In avoiding this high level of cognitive effort falsifiers show a higher level of satisficing behaviour by simply using one of the alternatives presented in the list. Similar behaviour (e.g. higher levels of satisficing) has been found when using filter questions (Hood and Bushery 1997). Filter questions allow a part of the questionnaire to be skipped. Thus, falsifiers use this option to reduce their cognitive effort and to save time.

However, falsifiers try to save time and reduce their cognitive effort only if legitimate opportunities are provided in the questionnaire. Otherwise falsifiers might show less satisficing than real respondents. This tendency is caused by the motivation of falsifiers to avoid detection. This is in line with findings indicating that falsifiers produce a lower level of item non-response than real respondents (Bredl et al. 2012; Shaeffer et al. 2005). Less item non-response on the part of falsifiers has been found for both open-ended and closed-ended questions. Since falsifiers tend to use only legitimate options in reducing their effort, at the same time they naturally avoid less legitimate options (e.g. item non-response). Falsified responses to open-ended questions require a level of effort which is probably still considered to be acceptable since open-ended questions are used quite rarely in standardized surveys. More frequent responding to open-ended questions seems to be inconsistent with less frequent usage of the "others" category in semi-open-ended questions. However, this may not be the case since falsifiers do not produce a high level of item non-response in the case of semi-open-ended questions, but rather a substantial response when they avoid "others, please specify" by ticking one of the available options.

A lower level of satisficing can also explain the previous finding that falsifiers produce less extreme responses than non-falsifiers (Shaeffer et al. 2005). Less extreme responding by falsifiers may also be associated with lower variances, as this has been found in falsified data by previous research (Shaeffer et al. 2005) – mainly for multi-item sets with rating scales. At first glance, lower variances (we refer to this as "non-differentiation") are not in line with the assumption of lower satisficing on the part of falsifiers. However, if we consider that usage of less extreme categories produces lower variances, we can understand higher non-differentiation as a secondary effect of less satisficing when providing less extreme responses. In addition, when using stereotypes falsifiers are not able to produce the same amount of variability in the data as that which is obtained in the real data. This can explain high non-differentiation in falsified data as well.

If we consider other response tendencies which are known for real respondents (Tourangeau et al. 2000) – middle responding, acquiescence, rounding, primacy and recency effects – we are able to generate additional hypotheses about formal indicators. In terms of all of these response tendencies we also assume less satisficing by falsifiers than by real respondents. In the following points we provide definitions and explanations for each of these response tendencies.

- Middle responding refers to the frequent usage of the middle category in rating scales; in doing so moderate opinions or neutral positions are indicated. We assume that falsifiers avoid middle responding so as to not make themselves conspicuous by producing only neutral opinions.
- Acquiescence is a tendency to agree, whereby real respondents tend to provide "Yes/Agree" responses regardless of item content (Messick 1976). This often occurs in multi-item sets with reversed wording of items. Here, respondents provide answers for negatively worded items in the same way that they answer positively worded items. In accordance with our assumption of less satisficing we suppose that falsifiers avoid this response pattern.
- Rounding occurs because respondents tend to answer open-ended questions about numerical information, for instance frequencies or quantities, using round numbers. Round numbers are multiples of 5, 10 or 7, the latter being used with questions related to the calendar (Tourangeau et al. 1997). Here a lower level of satisficing means that falsifiers answer metric questions with round numbers less frequently than real respondents.
- Primacy and recency effects are related to the order of presentation of response options (Tourangeau et al. 2000). These occur if sequences of response categories are presented to respondents, for example with "check all that apply" questions. A primacy effect might appear when response options are presented visually. In this case respondents prefer options located at the beginning of a list, meaning that they tend to choose the first option which seems satisfac-

tory, and they ignore the remaining options. This behaviour is a typical example of satisficing (Krosnick and Alwin 1987). Based on our assumption of a lower level of satisicing on the part of falsifiers, falsifers might produce a lower primacy effect.

- However, in the case of acoustic presentation of response alternatives, rather than primacy being an issue, there is instead a recency effect in real interviews. Recency effect means that respondents show a preference for choosing the last categories mentioned (Krosnick and Alwin 1987). Acoustic presentation is associated with the limited capacity of short-term memory, thus causing a recency effect. A falsifier uses a visual presentation of response categories (e.g. in a CAPI program), and thus a recency effect is less likely to influence falsified data. When generating indicators for primacy and recency effects it is important to use different presentation modes for lists of response categories.

In summary, we expect that falsifiers display reduced effort (higher satisficing) resulting in them choosing the "other, please specify" category less frequently and in them skipping subsequent questions in filter questions more frequently. In contrast, we expect less satisificing related to the remaining formal indicators, meaning that there is more effort on the part of falsifiers as compared to real respondents. This is expected to result in a lower level of item nonresponse, less frequent extreme and middle responding, acquiescence, rounding, and lower primacy and recency effects on the part of falsifiers. Next, we expect to find higher non-differentiation in falsified data resulting from avoidance of extreme categories and stereotyping on the part of falsifiers. In terms of content-related indicators in particular we suppose that questions with fictitious response categories are powerful in differentiating between real and falsified data. However, in the present study we used only relatively few content-related indicators since in our previous studies we found that formal indicators were associated with higher effects and more stable results than content-related indicators.

We have already tested some of the assumptions described in two preliminary studies (Menold, Storfinger and Winker 2011; Menold and Kemper 2011; Storfinger and Winker 2011). Both studies used subsets of data from the German General Social Survey (ALLBUS) 2008. Along with these subsets of real ALLBUS data, false data were produced by professional interviewers, students and researchers who played the role of falsifiers. These "falsifiers" were provided with some regional and socio-demographic information about real respondents. To test our assumptions concerning different indicators we then compared the real ALLBUS data and the data produced by the "falsifiers". In doing so we used an experimental design first applied by Hippler (1979). The results of these studies showed that “falsifiers” produced less extreme responses, less rounding, a lower recency effect and higher non-differentiation. Also, falsifiers used filter questions to skip some questions more often than real respondents did. In terms of responses to questions about behaviour we found that past political participation was considerably underestimated by falsifiers.

The aim of the current study is to validate the results of these studies and to test additional indicators (described above) which we were unable to test in the preliminary studies. In doing so we identify numerous indicators which we expect to be sensitive to falsifications, and we use them to improve the multivariate method proposed by Bredl et al. (2012).

Test of Indicators

Sample and Procedure

We used the approach by Hippler (1979), described above, for our explorative studies. The study was conducted in the summer of 2011. In order to compare real and falsified survey data as the first step we gathered N = 710 real face-to-face interviews. N = 78 students from the University of Giessen (age Md = 24 years, IQR = 3; 59.2% female, 97.9% single) were recruited as interviewers. The real respondents (age Md = 24 years, IQR = 3; 60% females, 97.6% single) were also students from the same university and they were recruited by the interviewers. On average each interviewer conducted 9.1 (SD = 1) interviews, mostly at respondents’ homes (58.1%) or on the university campus (28.4%). Each interview took about 30 minutes and each was audio recorded and checked for validity after the study was completed to ensure that interviews were actually conducted and not falsified. Interviewers received a payment of about eight Euros per interview. Interviewers were also provided with 20 Euros to serve as incentives which they could pay to respondents.

The second step was to have interviewers, who had just collected real data, fabricate survey data in the lab. In the process each interviewer received personal descriptions of real survey participants interviewed by his/her colleagues. These descriptions were randomly assigned to interviewers and they contained characteristics that a potential falsifier could obtain through a contact or short interview: e.g. sex, age, subject of study, number of semesters enrolled, marital status, residence, and country of origin. The interviewers now playing the role of falsifiers were briefed on the purpose of the study, which was to improve methods of identifying falsifications of survey data. They were also briefed on their role as falsifiers and on the procedure. The “falsifiers” were instructed to imagine conducting a face-to-face interview with the person described and then to fill in the survey questionnaire as they thought the person described would have responded to it. In this way they should try to provide answers that map the answers of the person described as closely as possible. The interviewers were paid three Euros for each falsified interview. We also provided a prize of 100 Euros for each of the three “best falsifiers”. An interviewer could win one of these prizes if their false data remained undetected following application of the multivariate detection method. In this manner we obtained a falsified data set of N = 710 falsified interviews corresponding to the N = 710 real interviews.

Questionnaire

The questionnaire used for collecting both real and false data contained 62 questions which were selected in such a way that they allowed for the construction of indicators sensitive to falsification. We predominantly used questions from the ALLBUS 2008. These questions covered different domains, such as attitudes toward political issues, attitudes toward women’s labour force participation, the economic situation, social justice, and political participation. Additionally, we used a questionnaire on personality (Big Five Inventory-10 (BFI-10; Rammstedt and John 2007) used in the International Social Program, ISSP). We also included a question related to magazine reading containing a portion of fictitious magazines (see above) and a measure of knowledge (Vocabulary and Overclaiming Test, VOCT; Ziegler et al. 2012). The VOCT includes terms which might be considered to be of common usage, as well as fictitious terms. The respondents were asked to assess which terms were known to them.

Construction of Indicators

Indicators were constructed based on some specific individual questions, as well as the multi-item measurements of attitude, behaviour, and personality mentioned above. In the case of multi-item measurements respondents indicated their responses on rating scales containing four to seven categories. Response scales were either unipolar or bipolar. In the current study we tested the following indicators:

1) SEMI-OPEN: relative frequencies of choosing the “others, please specify” category in four semi-open-ended questions included in the questionnaire (related to party vote, income sources, intended study degree and kind of study entrance diploma).
2) OPEN: in the questionnaire we used four open-ended questions related to an understanding of left and right orientation, magazines read and participation in sports. We counted the relative frequencies of providing responses to these open-ended questions to give us an indicator value.
3) FILTER: for this indicator we used ALLBUS questions pertaining to the father’s education and occupation when the respondents were 15 years old. These questions were asked in a subsequent step of a filter question. Here, choosing one of several response categories (e.g. father not known) allowed respondents to skip subsequent questions about their fathers. The frequency of choosing these categories was used as indicator.
4) INR (Item Non-Response): we counted the frequencies of item non-response across all questions (items) in the questionnaire (except for questions about the respondents' fathers and open-ended questions).
5) In terms of ERS (Extreme Responding Style) we counted the relative frequencies of choosing the most extreme responses on the rating scales, for example “1” or “5” on a five-point rating scale, across all items in multi-item measures (overall 42 items were considered).
6) For MRS (Middle Responding Style) we used only measures with an uneven number of response categories and we counted relative frequencies of choosing the middle category, again in the multi-item measures (in 26 variables).
7) ARS (Acquiescent Responding Style) was estimated based on BFI-10 responses (10 items). The BFI-10 contains five pairs of items which share a substantial amount of content but differ in terms of positive and negative item wording. ARS represents relative frequencies of agreement responses regardless of item direction.
8) ND (Non-Differentiation): we calculated the average standard deviation of responses across all items in multi-item sets (we used six multi-item sets).
9) ROUNDING: we calculated the relative frequency of rounded responses to numerical open questions. These included questions about the number of minutes spent watching television (rounded numbers here were: 30, 60, 90, 120, etc.), body mass information and income/payment information. For the latter students' income from different sources was queried: from parents, from study fundings, income from their own jobs.
10) PRIMACY: we counted how often (relative frequencies) the first two categories were chosen in a list of nominal response categories provided in four questions.

11) RECENCY: for two questions (about social class and social equity) interviewers read five to six categories in a rating scale to respondents. We then calculated how often respondents chose the last category.
12) NEWS: for this indicator we asked respondents what magazines they read and we provided them with a list of them, half of which were fictitious, as response categories. To create an indicator we calculated how often fictitious magazines were chosen (relative frequencies).
13) VOCT: We used the VOCT test and calculated the relative frequency of real words being correctly recognized as terms which actually exist.
14) PARTICIPATION: For this indicator past political activities were asked about using a "check all that apply" question. The relative frequency of political activities checked represents the value for this indicator.

Statistical Analysis

In order to test whether the indicators are sensitive to falsification as a first step we conducted a between-subjects multivariate analysis of variance (MANOVA) for false vs. real interviews. All of the indicators listed in the previous section were used as dependent variables. The MANOVA was conducted using SPSS 20. The values for dependent variables, with the exception of “non-differentiation” (ND), were not normally distributed in either of the treatment groups. However, in the case of large samples (n > 50) with equal group sizes, MANOVA results are quite robust to multivariate non-normality (see for instance Field 2009). Consequently, non-normality should not impact the results of MANOVA in our study since we used N = 710 for each of the treatment groups. The Box-Test pertaining to equality of covariance matrices can also be disregarded if equal group sizes are used (Field 2009). The Levene test only revealed equality of variances in both treatment groups for a few dependent variables. However, Hartley’s FMax test showed that the assumption of homogeneous variances within the groups is supported for all dependent variables.

Group membership for false and real data is known within the MANOVA. An alternative analysis is needed in order to find out how useful the indicators might be in detecting false data when group membership is not known. To this end we combined indicators which significantly differed between real and false interviews, as found by MANOVA, by means of multivariate statistical methods. A cluster analysis appears to be a promising way of using the multivariate structure of the indicator values to distinguish between the group of potential falsifiers and the honest interviewers (Bredl et al. 2012).

The central idea of a cluster analysis is to group similar elements together, while elements from different groups are naturally relatively heterogeneous. Due to the fact that we aimed to identify cheating interviewers and not simply single falsified interviews, the ideal outcome of the cluster analysis is the grouping of interviewers into two groups: one containing the falsifiers and the other containing the honest interviewers. Obviously based on a finite number of interviews per interviewer and many other factors affecting the distribution of indicator values such a perfect grouping cannot be expected in real applications. Nevertheless, this method did turn out to identify at least one cluster of susceptible interviewers in past survey applications, and this cluster included most if not all of the actual falsifiers.

To implement the clustering method each interviewer is characterized by a vector of d numerical values, whereby d corresponds to the number of indicators considered for the analysis. A single value is obtained for each indicator by pooling together all of the interviews done by one interviewer then calculating the values for the indicator based on this set of interviews. The aim of this cluster procedure was also to construct two clusters; one containing falsifiers and the other containing honest interviewers. Apart from the hierarchical clustering (Ward's method) used for the present analysis, we also tested the k-means approach. Since we obtained more reliable results using the hierarchical method as compared to k-means method, we only report the results based on Ward’s method here.

In addition to standard hierarchical clustering we also tested a clustering procedure which constructs clusters by global optimization of an objective function. The threshold acceptance (TA) algorithm (Winker 2001) was the method used within this framework to approximate the optimum solution. Starting with a randomly drawn assignment of the elements the TA procedure then assigns one element, also randomly drawn, of one cluster to the other cluster and accepts this new assignment as long as the modified cluster structure exhibits an improved value of the objective function or if it is at least not worse than the previous solution by more than a specific threshold. With this approach we allow for either an increase or a decrease in the quality of cluster structures in order to find the global optimum for the objective function under consideration, or at least a close approximation, after conducting a large number of local search steps. Based on heterogeneity within the clusters, measured by the sum of the pair wise distances in the respective cluster, the objective function used here should be minimized.

Obviously, given the assumptions regarding interviewer behaviour an optimal outcome for the cluster process is the formation of two clusters which can be separated exactly in terms of the indicators. To obtain such a result all interviewers in the “falsifier cluster” should, for example, show lower values for the proportion of “extreme” answers and for choosing the option “others”. This means that falsifiers should be separated from real respondents in exactly the manner which we supposed above. Given that such a perfect clustering is unlikely to occur in a real setting, an ex-post evaluation of the procedure is necessary to validate the approach. This is feasible in an experimental setting such as the one presented in this paper. In order to assess the performance of the method we can consider the proportion of interviewers who are correctly assigned as well as the proportion who are incorrectly classified. When doing this two types of errors can occur: type I errors, which involve a failure to assign a falsifier to the "falsifier cluster", and type II errors, which are "false alarms" indicating erroneously that an interviewer who performed all of his/her interviews correctly produced a falsification.

Results

Differences Between Real and Falsified Data

In this section we present the results regarding differences between real and falsified interviews. We calculated the MANOVAs on an individual case level and on an aggregated interviewer level. In this section we report the results calculated for each individual case since these do not differ from the results on an interviewer level. The results for the entire multivariate model show that there is at least one substantial difference between real and falsified data ($F_{(14,1405)} = 27.95$, $p < .001$, $\eta_p^2 = 0.22$). Next, we consider the univariate results concerning each indicator included in the model (Table 2.1).

Table 2.1: Univariate differences between real and falsified data regarding indicators of cheating

Indicator	Hypothesis	Con-firmed	Falsifica-tions M (SD)	Real Data M (SD)	$F_{(1,1418)}$	η_p^2
SEMI-OPEN	False < Real	yes	0.02 (0.07)	0.05 (0.11)	28.33***	0.02
FILTER	False > Real	yes	0.17 (0.38)	0.07 (0.27)	27.23***	0.02
OPEN	False > Real	no	0.94 (0.15)	0.97 (0.10)	21.79***	0.02
INR	False < Real	no	2.32 (2.65)	2.44 (2.21)	0.82	0.00
ERS	False < Real	yes	0.21 (0.13)	0.23 (0.11)	5.98*	0.004
MRS	False < Real	no	0.24 (0.12)	0.22 (0.10)	9.58**	0.007
ARS	False < Real	yes	0.46 (0.12)	0.51 (0.13)	65.48***	0.04
ND	False ≥ Real	yes	0.91 (0.18)	0.90 (0.16)	71.21***	0.05
ROUNDING	False < Real	no	0.56 (0.25)	0.45 (0.22)	88.72***	0.06
PRIMACY	False < Real	yes	0.33 (0.52)	0.43 (0.57)	12.21***	0.01
RECENCY	False < Real	yes	0.03 (0.13)	0.06 (0.16)	10.51**	0.007
NEWS	False > Real	yes	0.02 (0.07)	0.01 (0.04)	17.44***	0.01
VOCT	False ≤ Real	yes	0.34 (0.34)	0.44 (0.30)	32.92***	0.02
PARTICIPA-TION	False < Real	yes	0.37 (0.17)	0.43 (0.16)	51.94***	0.04

Note. *** p < .001; ** p < .01; * p < .05. Data for n = 710 real and n =710 falsified interviews. For explanation of indicators see “Construction of Indicators”.

As described above we expected to see a limited effort on the part of falsifiers as compared with real respondents, in particular through less frequent usage of the category “others” in semi-open-ended questions and in more frequent responses to filter questions which allow skipping of subsequent questions. As can be seen in Table 2.1, there are significant differences with respect to semi-open-ended and filter questions, and these are in line with our expectations.

For another set of indicators we expected more careful response behaviour – in other words less satisficing – by falsifiers than by real respondents. These indicators were: frequency of answering open-ended questions (OPEN), item non-response (INR), extreme responding (ERS), middle responding (MRS), acquiescent responding (ARS), non-differentiation (ND), rounding, primacy and recency effects. The results support a majority of our hypotheses as they show less extreme responding, less acquiescence, higher non-differentiation, as well as lower primacy and recency effects for falsified data as compared with real data. All of these differences are significant. The results for item non-response (INR) showed fewer unanswered items on the part of the falsifiers, although this result is not significant.

However, there are three significant differences which are not in line with our assumption of less satisficing on the part of falsifiers. In fact, falsifiers demonstrated a higher level of satisficing than real respondents since they provided fewer responses to open-ended questions, more rounded answers to openly asked numerical questions (indicator ROUNDING) and a higher level of middle responding style (MRS).

Next, we confirmed our assumptions concerning a higher selection of fictitious response categories by falsifiers. In fact, falsifiers chose fictitious magazines more often to describe the reading behaviour of respondents (NEWS). Similarly, falsifiers also used fictitious terms more often to describe the vocabulary known to respondents (VOCT). Furthermore, past political participation was strongly underestimated by falsifiers, as has also been shown by our previous explorative research.

In summary, we found that all indicators produced significant differences between real and false data, with the exception of item non-response. Thus, these can be used as indicators of cheating behaviour in the subsequent cluster analysis. The results also show that the assumption about "less satisficing" applied to the majority of respective indicators. However, in the case of three indicators reduced effort by falsifiers was observed. We assume that in this case falsifiers' motivation to remain undetected (which we expected to result in less satisficing) conflicts with their motivation to save time. This is likely since in our experimental design the consequences related to being detected were not nearly as serious as those in a real survey context. In addition we explain the result related to open-ended questions by an excessively high level of effort required from falsifiers in the current study: we included not just one but rather a number of open-ended questions which had to be answered for approximately 10 respondents. To explain the result concerning ROUNDING we compared the current results with results from our previous explorative studies. In the current study we asked for specific individual income information and mainly on this basis we calculated a value for the rounding indicator. In previous studies (e.g. Menold et al. 2011) we used a classic total net household income question which has been commonly used in surveys. This question requires consideration of each person living in the household and all income sources, which may be associated with memory gaps on the part of real respondents, who then round numbers more often than falsifiers do. Therefore, we conclude that availability of exact information on the part of real respondents could explain contradictory results found in our different studies. If exact information is not available in the memory of real respondents (e.g. when asking about household net income) rounding should become more relevant in real data than in falsified data. In contrast, if exact numerical information is known to respondents (e.g. individual incomes) then non-falsifiers will demonstrate less rounding than falsifiers. A future subject of research will be to study in more detail the impacts of the likelihood of detection, task difficulty and information availability on falsifiers' behaviour.

Cluster Analysis for Separating Real and False Data

Based on the significant results of the univariate analysis in terms of differences in indicator values between falsified and real data (see Table 2.1) we decided to consider 13 indicators for the cluster analysis at the interviewer level. The nonresponse indicator (INR) was omitted due to insignificant results in the univariate analysis (see Table 2.1).

As a first step, we provide the results regarding the cluster performance of each indicator separately. In the second step, we report the results for the multivariate cluster analysis including all 13 indicators simultaneously. The effectiveness of the analysis is always judged based on the proportion of correctly assigned falsifiers as well as the proportion of correctly assigned honest interviewers (non-falsifiers).

As mentioned above we aimed to identify falsifiers and not single falsified interviews. Consequently, this analysis required calculation of the 13 indicator values on an interviewer level. Hence, we had to bring together the indicator values by pooling together all interviews conducted by one interviewer and then we had to calculate the respective indicator value. This was repeated for each of the 156 interviewers (78 falsifiers and 78 non-falsifiers) in the dataset. As reported in Table 2.1 calculating indicator values at the interviewer level produces similar mean values and significant differences between falsifiers and non-falsifiers. But we can also note that the variance of all indicator values is higher within the cheating group than within the group of non-falsifiers.

Table 2.2 shows the results of the hierarchical clustering method. The proportion of falsifiers who were correctly identified ranges between 17% and 91%, while the proportion of correctly classified non-falsifiers varies between 26% and 100%.

Table 2.2: Cluster results based on the hierarchical clustering method (Data: 78 falsifiers and 78 non-falsifiers)

	CLUSTER A (Falsifiers)			**CLUSTER B (Non-Falsifiers)**		
Indicator	**Cluster-size**	**Falsifiers** (% of all falsifiers)	**Non-Falsi-fiers**	**Cluster-size**	**Non-Falsifiers** (% of all non-falsifiers)	**Falsi-fiers**
SEMI-OPEN	129	71 (91%)	58	27	20 (26%)	7
FILTER	43	32 (41%)	11	113	67 (86%)	46
OPEN	39	29 (37%)	10	117	68 (87%)	49
ERS	33	29 (37%)	4	123	74 (95%)	49
MRS	102	51 (65%)	51	54	27 (35%)	27
ARS	83	57 (73%)	26	73	52 (67%)	21
ROUNDING	21	21 (27%)	0	135	78 (100%)	57
PRIMACY	55	36 (46%)	19	101	59 (76%)	42
ND	86	51 (65%)	35	70	43 (55%)	27
RECENCY	127	70 (90%)	57	29	21 (27%)	8
NEWS	15	13 (17%)	2	141	76 (97%)	65
VOCT	31	29 (37%)	2	125	76 (97%)	49
PARTICIP.	35	27 (35%)	8	121	70 (90%)	51
ALL 13 INDICAT.	**44**	**43 (55%)**	**1**	**112**	**77 (99%)**	**35**
Subset 1	23	23 (29%)	0	133	78 (100%)	55
Subset 2	145	74 (95%)	71	11	7 (0.1%)	4
Subset 3	**79**	**57 (73%)**	**22**	**77**	**56 (72%)**	**21**

When comparing results for the 13 indicators we can see that acquiescence (ARS) and non-differentiation (ND) are indicators which yield outstanding results in terms of identifying real and falsified data. When using acquiescence as an indicator we found that 73% of falsifiers and 67% of non-falsifiers were correctly identified. When using the ND indicator 55% of non-falsifiers and 65% of falsifiers were revealed.

With respect to success only in terms of identifying falsifiers, Table 2.2 shows that nearly all falsifiers could be identified using the RECENCY indicator (90%) as well as the SEMI-OPEN indicator (91%). However, this result is not satisfactory because these two indicators produce the largest “falsifier-clusters” (see Table 2.2). More than two thirds of all interviewers are assigned to this cluster and consequently a high proportion of false alarms is produced. This is also true of the midpoint-ratio (MRS), which reveals more than half of the falsifiers (65%) but the obtained falsifier-cluster itself contains approximately 65% of all interviewers. All other indicators produced fewer false positives but they were only able to identify less than half of the falsifiers when they were considered individually (the proportion varies between 17% and 46%).

While the results related to identifying falsifiers using individual indicators are only satisfying in some cases, identifying non-falsifiers was somewhat more successful. Eight of the 13 indicators detected more than three quarters of all non-falsifiers (proportions range between 76% and 100%) but at the same time high proportions of false negatives are produced. All other indicators were able to identify less than 70% of the non-falsifiers.

As proposed by Bredl et al. (2012), the proportion of correctly assigned interviewers may increase if all 13 indicators are used simultaneously in a multivariate analysis. This way not only the identifying power of each indicator is exploited, but also the dependency structures between indicators, which could increase the overall quality of assignment into the correct group. In fact the results for our dataset are quite satisfying (see Table 2.2) given that assignment to the "non-falsifier-cluster" is successful in almost all cases (99%) and slightly more than half of the falsifiers were assigned to the falsifiers' cluster (55%).

To conclude whether the multivariate approach performs better than the 13 single analyses we must consider the precision of the cluster result, for example by investigating cluster sizes in relation to the proportion of correctly assigned interviewers. From the point of view of practitioners in polling agencies we prefer highly precise values, especially for the "falsifier-cluster". If polling agencies use the cluster result for further checks, such as re-interviews, the number of non-falsifiers in the "falsifier-cluster" should be as small as possible, and of course the number of falsifiers should be as high as possible. Thus, polling agencies could avoid inspecting a high proportion of false alarms. When considering the precise values for the 13 single analyses, we can notice that a few indicators produce very small proportions of false alarms (e.g. ROUNDING and NEWS) but at the same time less than half of the falsifiers are assigned to the respective "falsifier-cluster". Only the multivariate approach provides a satisfactory proportion of correctly assigned falsifiers and at the same time a very small proportion of false alarms. Hence, we recommend relying on the multivariate approach.

Nevertheless, we were interested in increasing the proportion of correctly assigned falsifiers. In doing so we considered smaller subsets of indicators to investigate how the results change. The selection of indicators to be used simultaneously was based on the results obtained from the 13 single (univariate) analyses. Firstly we decided to incorporate only the acquiescence indicator (ARS) and the non-differentiation value (ND) because of their outstanding results in the univariate analysis (Subset 1). Secondly we investigated the SEMI-OPEN and RECENCY indicators because they delivered the highest values in terms of identifying falsifiers in the univariate analyses (Subset 2). For the last subset of indicators we chose all indicators which performed better than the multivariate approach (using all 13 indicators), either in identifying falsifiers (SEMI-OPEN, MRS, ARS, ND and RECENCY) or in identifying non-falsifiers (ROUNDING) (Subset 3).

Table 2.2 shows the results for the three different subsets. The first two subsets succeed only in increasing either the proportion of identified falsifiers or the proportion of correctly assigned non-falsifiers. However the third subset provided high values for both shares, even if the number of false alarms in the "falsifier-cluster" was higher than the value obtained when using all 13 indicators. Consequently, the number of indicators used for the hierarchical cluster analysis depends on the specific interest. When the aim is to produce a precise "falsifier-cluster", for example with low false alarm rates, all 13 indicators should be used. But if we are interested in correctly assigning a high share of all interviewers we might use a smaller subset of indicators.

In addition to hierarchical clustering we also applied a global clustering method based on heuristic optimization (TA) to our data set. In terms of the results for the 13 analyses using only one indicator separately (see Table 2.3) we found that using the global clustering method produces high proportions of correctly assigned falsifiers. Ten of the 13 univariate analyses identified more than half of the 78 falsifiers in the data set (proportions vary between 54% and 83%). Identifying more than half of the non-falsifiers was successful in all 13 single analyses. Also the multivariate approach when using 13 indicators leads to very satisfying results. It was possible to correctly assign 82% of falsifiers and 92% of non-falsifiers.

If we compare these results with those obtained using hierarchical clustering we can notice that the global clustering method provides equal cluster sizes in the majority of the cases. This may be advantageous for our analysis because our dataset contains 50% falsifiers and consequently the proportions of falsifiers revealed are almost always higher than those when using the hierarchical clustering method. Hence, we recommend using the global clustering method instead of the hierarchical approach when the aim is precise differentiation between falsifiers and non-falsifiers. The global clustering method using all 13 indicators simultaneously leads to even higher proportions of correctly assigned interviewers than the hierarchical approach using the small subset of indicators. If we take a look at the results of the TA approach using the three subsets (see Table 2.3) we can notice that the proportions of correctly assigned falsifiers and non-falsifiers are slightly less convincing when compared to the multivariate analysis using all 13 indicators.

Table 2.3: Cluster result based on the global clustering method (TA) (Data: 78 falsifiers and 78 non-falsifiers)

	CLUSTER A (Falsifiers)			**CLUSTER B (Non-Falsifiers)**		
Indicator	**Cluster-size**	**Falsifiers** (% of all falsifiers)	**Non-Falsifiers**	**Cluster-size**	**Non-Falsifiers** (% of all non-falsifers)	**Falsifiers**
SEMI-OPEN	105	65 (83%)	40	51	38 (49%)	13
FILTER	47	33 (42%)	14	109	64 (82%)	45
OPEN	58	37 (47%)	21	98	57 (73%)	41
ERS	81	43 (55%)	38	75	40 (51%)	35
MRS	72	42 (54%)	30	84	48 (62%)	36
ARS	83	57 (73%)	26	73	52 (67%)	21
ROUNDING	72	55 (71%)	17	84	61 (72%)	23
PRIMACY	82	48 (62%)	34	74	44 (56%)	30
ND	69	46 (59%)	23	87	55 (71%)	32
RECENCY	94	56 (72%)	38	62	40 (51%)	22
NEWS	40	28 (36%)	12	116	66 (85%)	50
VOCT	71	43 (55%)	28	85	50 (64%)	35
PARTICIP.	74	50 (64%)	24	82	54 (69%)	28
ALL 13 INDICAT.	**70**	**64 (82%)**	**6**	**86**	**72 (92%)**	**14**
Subset 1	77	56 (72%)	21	79	57 (73%)	22
Subset 2	97	61 (78%)	36	59	42 (54%)	17
Subset 3	**77**	**58 (74%)**	**19**	**79**	**59 (76%)**	**20**

In summary applying the hierarchical clustering method is inferior to using the global clustering method (TA). Using Ward’s method often results in low proportions of false alarms, but this comes at the cost of a high proportion of overlooked falsifiers. Hence, we recommend using the global clustering procedure (TA) in order to achieve a relatively high proportion of revealed falsifiers.

Conclusions

The aim of the study presented in this article was to test numerous indicators in terms of their effectiveness in helping to identify falsified data in face-to-face surveys. The main idea behind this approach was to use differences between real and falsified data as a source of indicators, and to combine these within a multivariate cluster analysis.

The effectiveness of indicators was expected to result from differences in cognitions and motivation between real respondents and falsifiers. We were able to confirm the majority of our assumptions concerning these differences.

In addition, nearly all of the indicators tested are useful for identifying falsified data. Using a selected set of indicators leads to the best performance for the clustering method. This shows that some indicators are more powerful than others. We were able to identify most of the falsifiers in the large data set produced in our experimental setting. However, it was not possible to obtain convincing results with a standard clustering method (e.g. hierarchical cluster). On the other hand an optimised clustering method (TA) was more effective. Further research is needed to verify the extent to which this result can be generalized to other surveys and contexts. In real settings we assume that the proportion of falsifiers is much smaller, ranging perhaps between 5 and 8 % (Schnell 1991). In such a setting identifying falsifications based on the clustering method might be even more successful (Storfinger and Winker 2011). To test this assumption we created synthetic data by means of bootstrapping. This implies that the number of falsifiers is modified artificially by re-sampling individual interviews of randomly selected interviewers. More details on this procedure and the obtained results are provided by Storfinger and Winker (2013) in this volume.

With regard to the number of clusters created in our identification method further developments should be considered. Thus, it could be helpful to create three (instead of two) clusters. In doing so falsifiers themselves could be clustered into two groups: one cluster containing interviewers who falsified data in the way that we have supposed (see hypotheses) and the second cluster containing falsifiers who falsified employ a different strategy. For instance, we found that typical falsifiers avoid extreme answers (ERS), but the dataset shows that there are also some falsifiers who prefer extreme answers even more than non-falsifiers do. First analyses showed that the cluster result improves significantly if we calculate three clusters, but this is only true for a few indicators. Further research on the motivation of falsifiers is necessary in order to understand the relationship between indicator values and the number of clusters.

Despite these remarks regarding possible further developments, it also appears to be important to test the detection method presented in this paper in a real survey setting. This method can be used in surveys as a supplement to other control procedures during data collection. Identifying interviewers who are “at risk” by using this multivariate approach would be more productive for subsequent extended controls compared with single indicators, a method used in current survey practice (e.g. differences in sex and age, as described by Koch 1995). In this way it would be possible to obtain not only higher quality data but also higher validity data regarding the prevalence of falsifications in surveys.

Bibliography

Biemer, Paul. P., Stokes, S. Lynne.: The optimal design of quality control samples to detect interviewer cheating, Journal of Official Statistics, 5(1), 23-39, 1989.

Blasius, Jörg, Thiessen, Victor: Detecting poorly conducted interviews. In Winker, Peter, Menold, Natalja and Porst, Rolf (Eds.): Survey Standardization and Interviewer’s Deviations – Impact, Reasons, Detection and Prevention. Peter Lang, Frankfurt a. M., 2013, pp. 67 – 88.

Bredl, Sebastian, Winker, Peter, Kötschau, Kerstin: A statistical approach to detect interviewer falsification of survey data, Survey Methodology, 38(1), 1-10, 2012.

Field, Andy: Discovering statistics using SPSS, Singapore, Sage, 2009.

Groves, Robert M., Fowler, Floyd Jackson Jr., Couper, Mick P., Lepkowski, James M., Singer, Eleanor, Tourangeau, Roger: Survey Methodology, New York, Wiley, 2004.

Güllner, Gesine, Porst, Rolf: Identifikation von Fälschungen in Surveys. Bericht über den kognitiven Test im Rahmen des Projekts IFiS, GESIS-Working Papers, 16, (2012).

Hippler, Hans-Jürgen: Untersuchung zur "Qualität" absichtlich gefälschter Interviews, ZUMA-Arbeitspapier, Mannheim, 1979.

Hood, Catherine C., Bushery, John M.: Getting more bang from the reinterview buck, Identifying "at risk" interviewers, Proceedings of the American Statistical Association, Survey Research Methods Section 27, 820-824, 1997.

Koch, Achim.: Gefälschte Interviews, Ergebnisse der Interviewerkontrolle beim ALLBUS 1994, ZUMA-Nachrichten, 36, 89-105, 1995.

Krosnick, Jon A., Alwin, Duane F.: An evaluation of a cognitive theory of response-order effects in survey measurement, Public Opinion Quarterly, 51, 201-219, 1987.

Menold, Natalja, Kemper, Christoph J.: 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, Natalja, Storfinger, Nina, Winker, Peter.: Development of a method for ex-post identificationof falsifications in survey data, Proceedings of New Techniques and Technologies for Statistics - NTTS 2011, Brussels, Belgium, 2011.

Messick, Samuel J.: The psychology of acquiescence, An interpretation of research evidence, In: Berg, I. A. (Ed.), Response Set in Personality Assessment, Chicago, Aldine Publishing Company, 1967.

Noelle-Neumann, Elisabeth: Ein Wissenschafter muss Neuland betreten, Wiener Zeitung, April, 17., 2003 (http://www.philosophische-praxis.at/noelle.html, August 30 2012)

Rammstedt, Beatrice, John, Oliver P.: Measuring personality in one minute or less, A 10-item short version of the Big Five Inventory in English and German, Journal of Research in Personality, 41(1), 203-212, 2007.

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

Schäfer, Christin, Schräpler, Jörg-Peter, Müller, Klaus-Robert, Wagner, Gert G.: Automatic identification of faked and fraudulent interviews in the German SOEP, Journal of Applied Social Science (Schmollers Jahrbuch), 125 (1), 183-193, 2005.

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

Schräpler, Jörg-Peter, Wagner, Gert G.: 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, 2003.

Schreiner, Irwin, Pennie, Karen, Newbrough, Jennifer: Interviewer falsification in census bureau surveys, Proceedings of the American Statistical Association, Survey Research Methods Section XII, 491-496, 1988.

Shaeffer, Eric M., Krosnick, Jon A., Langer, Gary E., Merkle, Daniel M.: Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions, Public Opinion Quarterly, 69(3), 417-428, 2005.

Storfinger, Nina, Winker, Peter: Robustness of clustering methods for identification of potential falsifications in survey data, ZEU Discussion Paper 57, Giessen, 2011.

Storfinger, Nina, Winker, Peter: Assessing the performance of clustering methods in falsification using bootstrap. In Winker, Peter, Menold, Natalja and Porst, Rolf (Eds.): Survey Standardization and Interviewer's Deviations – Impact, Reasons, Detection and Prevention. Peter Lang, Frankfurt a. M., 2013, pp. 49-65.

Stroebe, Wolfgang, Hewstone, Miles, Codol, Jean-Paul, Stephenson, Geioffrey: Sozialpsycholgie. Eine Einführung. Berlin et al.: Springer, 1992.

Tajfel, Henri. Cognitive aspects of prejustice, Journal of Social Issues, 25, 79-97, 1969.

Tourangeau, Roger, Rasinski, Kenneth A., Jobe, Jared B., Smith, Tom W., Pratt, William F.: Sources of error in a survey on sexual behavior, Journal of Official Statistics, 13(4), 341-365, 1997.

Tourangeau, Roger, Rips, Lance J., Rasinski, Kenneth A.: The psychology of survey response, Cambridge University Press, 2000.

Wetzel, Angela-Jo: Assessing the effect of different instrument modes on reinterview results from the consumer expenditure quarterly interview survey, Proceedings of the American Statistical Association, Survey Research Methods Section 435, 4508-4513, 2003.

Winker, Peter: Optimization heuristics in econometrics, Applications of threshold accepting, Chichester, Wiley, 2001.

Ziegler, Matthias, Kemper, Christoph J., Rammstedt, Beatrice: The vocabulary and overclaiming test (VOC-T), Manuscript submitted for publication, 2012.