DE GRUYTER OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 643–660, http://dx.doi.org/10.1515/JOS-2016-0033

Detecting Fraudulent Interviewers by Improved Clustering Methods – The Case of Falsifications of Answers to Parts of a Questionnaire

Samuel De Haas[1] *and Peter Winker*[1]

Falsified interviews represent a serious threat to empirical research based on survey data. The identification of such cases is important to ensure data quality. Applying cluster analysis to a set of indicators helps to identify suspicious interviewers when a substantial share of all of their interviews are complete falsifications, as shown by previous research. This analysis is extended to the case when only a share of questions within all interviews provided by an interviewer is fabricated. The assessment is based on synthetic datasets with a priori set properties. These are constructed from a unique experimental dataset containing both real and fabricated data for each respondent. Such a bootstrap approach makes it possible to evaluate the robustness of the method when the share of fabricated answers per interview decreases. The results indicate a substantial loss of discriminatory power in the standard cluster analysis if the share of fabricated answers within an interview becomes small. Using a novel cluster method which allows imposing constraints on cluster sizes, performance can be improved, in particular when only few falsifiers are present. This new approach will help to increase the robustness of survey data by detecting potential falsifiers more reliably.

Key words: Survey data falsifications; partial falsifications; cluster analysis; constraint cluster analysis; bootstrap.

1. Introduction

Survey data are a central ingredient of empirical research in economics, other social sciences, and medicine. The quality of any analysis of surveys depends on the quality of the survey data. A huge literature exists on potential pitfalls linked to issues of sampling and the construction of questionnaires which might have a negative impact on data quality. The issue of potential falsifications by the interviewers, however, has received less attention, although anecdotal reports date back to Crespi (1945). In fact, the prevalence of such behavior might be higher than commonly assumed. For a recent survey of the literature, see Bredl et al. (2013). They conclude that the share might be typically below five percent for large scale surveys with intensive supervision, while it might reach levels exceeding 50% in smaller surveys with limited supervision and difficult framework conditions such as inaccessibility of respondents or binding quota requirements.

[1] University of Giessen, Chair of Industrial Organisation, Regulation and Antitrust, and Chair of Statistics and Econometrics, Licher Strasse 64, 35394 Giessen, Germany. Emails: Samuel.De-Haas@wirtschaft.uni-giessen.de and Peter.Winker@wirtschaft.uni-giessen.de

Acknowledgments: Financial support through the DFG in project WI 2024/5-4 is gratefully acknowledged. We are indebted to the associate editor handling our submission and three anonymous referees for their comments which helped to improve the presentation of our results substantially.

Consequently, approaches focusing both on prevention and deterrence are required. Concerning prevention, this might include appropriate interviewer training, payment, and motivation (Gwartney 2013). Approaches for deterrence include close supervision and controls in the field and of the collected data.

Bredl et al. (2012) proposed a method for the analysis of collected data, which employs a clustering procedure on multivariate indicators calculated at interviewer level. The method has been tested successfully on real and experimental datasets (Menold et al. 2013). The method was not meant to replace other methods used for quality management such as reinterviews (Forsman and Schreiner 1991), but rather to focus them on a subset of interviewers exhibiting conspicuous patterns in the data they contribute. The number of accessible real and experimental datasets with identified falsifications is limited, as there are no incentives to report such cases in real surveys (for an exception see Finn and Ranchhod 2013). Although reporting identified fabrications in survey data might be considered a clear signal of successful supervision and quality management, it might also mislead the reader into challenging the integrity of the data. For this reason, identified falsifications in survey data are typically removed before the data are made available for further research without explicit indication. As a consequence, the robustness of the method with regard to the choice of indicators and the structure of the dataset (number of interviewers, share of falsifiers) cannot be assessed solely by using the few datasets available. Insights into this problem can be gained by generating synthetic data from real or experimental data using a bootstrap-based approach as described by Storfinger and Winker (2013).

The bootstrap approach has been used to analyze the performance of the clustering procedure for partial falsifications, that is, the situation when interviewers provide some real interviews and add falsifications, for example, to complete a quota (De Haas and Winker 2014). The present article complements this earlier work with an analysis of partial falsifications within questionnaires, that is, for the situation when interviewers collect part of the data from the respondents and complete the questionnaire themselves afterwards. Anecdotal evidence suggests that there are different reasons for both types of partial falsifications in real surveys. Examples are that part of the questionnaire comprises embarrassing questions that the respondents refuse to answer or questions which are time-consuming when filled in with the respondent. As in previous work (De Haas and Winker 2014), we are interested in the effects of shrinking shares of fabricated answers by a deviant interviewer on the performance of the clustering procedure.

We add a novel clustering tool for the present analysis, which allows us to impose a priori constraints on the (expected) maximum number of falsifiers. This approach is motivated by the finding that the unconstrained clustering method tends to produce a substantial share of false alarms, especially when the share of falsifiers is low or only partial falsifications are provided (De Haas and Winker 2014). Using the constrained approach might improve the discriminatory power of the method, in particular, when the share of falsifiers is small. Finally, we add Matthew's correlation coefficient (Matthews 1975) as an alternative summary measure. It complements the standard measures used for the quality of the assignment of interviewers to the two groups of honets and supposedly deviant interviewers, namely oversights and false alarms.

The article is organized as follows. In Section 2 we introduce the methods used for the identification of falsifications, in particular the indicators constructed at the interviewer

level, the standard clustering procedure, and the new variant imposing a size constraint on the falsifier cluster. The experiment providing the data is described in Section 3 together with the bootstrap procedure for generating synthetic data with a specific structure. The results are summarized in Section 4, while Section 5 concludes and provides an outlook onto further steps of the research on partial falsifications.

2. Methods

The data-based identification of potential falsifications uses properties of the data which differ between real and falsified interviews. The indicators used for that purpose should ideally be independent of a specific questionnaire. At the same time, they should be unknown to the interviewers to avoid strategic falsifications. Several indicators have been proposed and used previously (Schäfer et al. 2005; Kemper et al. 2011; Storfinger and Opper 2011; Bredl et al. 2012; Menold et al. 2013; Kemper and Menold 2014; Menold and Kemper 2014; De Haas and Winker 2014). To allow for a comparison of the results, we use the same indicators as De Haas and Winker (2014). A full list of these indicators with a short description is provided in Appendix A.

In the following, the indicator acquiescent responding style (ARS), which has been used previously by Kemper et al. (2011), illustrates the idea of using indicators to separate real and fabricated interviews. It is constructed based on pairs of items which address similar issues, but differ in using either a positive or negative wording, respectively. Consequently, fully rational respondents should choose opposite answers. However, it is commonly observed that respondents tend to prefer to agree with a given statement and to some extent provide inconsistent answers to such pairs of questions. While some interviewers might be aware of these phenomena, it may be impossible for them to judge the extent of such acquiescent behavior by real respondents. In fact, results from previous studies show that falsifiers tend to exhibit less acquiescence in their fabricated interviews (Menold et al. 2013). The indicator ARS used for the present application is based on five pairs of such items and measures the relative agreement frequency. Here, agreement frequency is defined as the share of the answer options "fully correct" and "fairly correct".

While most indicators can be calculated for each questionnaire, it appears doubtful that they would allow for a discrimination at the level of individual interviews. Typically, the number of questions linked to each indicator is rather small unless the questionnaires are extensively long. For this reason, as in previous research (Menold et al. 2013; De Haas and Winker 2014), we focus our analysis on the interviewer level. Thus, the values of all indicators are calculated based on all interviews for each interviewer. We will consider an interviewer as deviant ("falsifier") if at least one of her or his interviews or parts thereof are not obtained from real respondents, but are fabricated by the interviewer her- or himself. Obviously, given the aggregation at the interviewer level, detection might become more difficult if the share of fabricated (parts of) interviews is low.

Differences between real and false interviews might show up simultaneously in several indicators. If these indicators are not perfectly correlated (for most of the indicators used here, the pairwise correlation is found to be smaller than 0.2), exploiting the multivariate structure in a cluster analysis is expected to outperform splits based on a single indicator. This idea is supported by the findings presented in earlier research (Bredl et al. 2012;

Menold et al. 2013). Furthermore, using the multivariate distribution of several indicators makes it more difficult for a deviant interviewer to generate data meeting the properties of real data closely enough to pass through undetected.

For the cluster analysis, each interviewer k is represented by a vector of indicator values $\mathbf{i}_{k,j}, k = 1, \ldots, K$ and $j = 1, \ldots, J$. K denotes the total number of interviewers and J the number of indicators. Prior to performing the cluster analysis each indicator value is standardized, resulting in

$$\tilde{\mathbf{i}}_{k,j} = \frac{\mathbf{i}_{k,j} - \bar{\mathbf{i}}_{.,j}}{\sqrt{\text{var}(\mathbf{i}_{.,j})}}.$$

Often, the task of clustering a set of vectors like $\tilde{\mathbf{i}}_k = (\tilde{\mathbf{i}}_{k,1}, \ldots, \tilde{\mathbf{i}}_{k,J})$ is tackled by using hierarchical (agglomerative or divisive) clustering methods (Baragona et al. 2011, 199ff). However, the sequential approach of agglomerative procedures might not result in a global optimum for the assignment. Consequently, we apply a global clustering approach. While existing methods such as k-means (Baragona et al. 2011, 211) could be employed in this case, they also do not guarantee convergence to a global optimum given their iterative local search. Here, we use Threshold Accepting (TA) because it is a globally convergent optimization heuristic. It has been shown that under certain conditions it will converge to the global optimum when the number of search steps goes to infinity (Althöfer and Koschnik 1991; Winker 2001). Other optimization heuristics might also be used in this context, for example the clustering method based on genetic algorithms described by Baragona et al. (2011, 219ff). A further advantage of using this approach is the possibility to add constraints. In our application, this option will be used to limit the size of the cluster corresponding to potential falsifications.

We start with the description of the unconstrained version of the algorithm as used previously by Storfinger and Winker (2013) and De Haas and Winker (2014): it aims at minimizing an objective function which is calculated as the sum of the pairwise Euclidean distances within the clusters. Hence, the goal consists in reducing the heterogeneity with regard to the values of the various indicators within each group. The algorithm is initialized with a randomly drawn assignment of all elements (interviewers) into two groups. Afterwards, for a preset number of iterations, one randomly chosen element is regrouped in each iteration. The resulting new clusters are accepted as long as the value of the objective function is improved (decreases) or at least does not increase by more than a predefined threshold. In order to find a global optimum, or at least a close approximation, this local search step has to be repeated many times. An obvious drawback of the optimization-based clustering method as compared to traditional clustering algorithms is its higher computational cost. On a standard desktop computer (i7-3770 CPU, 3.40 GHz, 8 GB RAM) a single run of the Matlab implementation with 2,500,000 iterations takes about 110 seconds to finish. This becomes relevant in a simulation study as the present one, when the clustering problem has to be solved thousands of times, while it is not a major issue for a single application to a single real dataset in a survey setting.

After the clustering algorithm is completed, the identification of the two clusters is based on the assumptions about the indicator values for honest versus cheating interviewers (see Table 2 in Appendix A). Therefore, it can be decided automatically

which of the two subgroups represents the falsifiers and the honest interviewers, respectively. In order to perform this identification step unsupervised for each bootstrap simulation, we sum up the standardized mean values for every indicator over all interviewers in each cluster. To this end the signs of the indicators are adjusted in such a way that higher values always point to the group containing the potential falsifiers.

Employing this clustering procedure may result in two potential types of errors. Honest interviewers might be incorrectly added to the group labeled "deviant interviewers". Such a misassignment is called "false alarm" or "false negative". On the other hand, some falsifiers might be allocated to the group labeled "honest interviewers". This type of error is called "oversight" or "false positive". To provide a summary measure of the extent of such misclassifications, Matthews's correlation coefficient (MCC) is used (Matthews 1975). The MCC is calculated as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}, \tag{1}$$

where TP denotes the number of true positives, that is, the number of interviewers correctly assigned to the group labeled as "honest interviewers". TN is the number of true negatives, that is, of correctly identified falsifiers. FP is the number of false positives (oversights) and FN the number of false negatives (false alarms) (for a comprehensive overview of alternative measures used to evaluate the quality of binary classifications see also Verbiest et al. 2015). By construction, MCC takes on values between -1 and 1. If all interviewers are correctly assigned, it takes on the value one, for a random, that is, noninformative assignment, the values should be close to zero, while when exactly the wrong assignment is given, MCC takes on the value -1.

The objective function used lays highest emphasis on the homogeneity within its clusters based on pairwise Euclidian distances. Implicitly, this objective functions favors clusterings with rather similar group sizes. Therefore, it might be expected to perform better if the share of falsifiers is about 50%, while it tends to generate a large number of false alarms for a low share of falsifiers in the dataset. Consequently, imposing some additional constraints on the size of the falsifier cluster might be beneficial, in particular if a modest share of falsifications is expected and the cost of controls is high. It might help to concentrate controls on those interviewers with the highest risk. Obviously, if some falsifications are found in this group, there is a risk that even a larger share of interviewers is actually deviating and the analysis might need to be repeated allowing for a larger size of the falsifier cluster.

Technically, the constraint is imposed in the following way: the same optimization heuristic as described above is used. The objective function is augmented by an additional term, described as "penalty term". This part of the objective function takes on the value zero as long as the number of interviewers assigned to the potential falsifiers cluster is smaller than or equal to the predefined limit. If more interviewers are assigned to this cluster, the term becomes positive and is an increasing function of the differences between the number of potential falsifiers and the predefined value. Due to this penalty term, the value of the objective function decreases when the falsifier cluster becomes smaller as long as its size is above the predefined limit. Consequently, as the algorithm proceeds and the weight of the penalty increases, the current solution becomes more and more likely to

satisfy the imposed constraint, which should be met by the final solution. For a description of alternative ways to handle constraints in the framework of an optimization heuristic, see Gilli et al. (2011, 352ff). The computational complexity of the algorithm is higher when the constraint is taken into account, as the evaluation of the penalty term requires the identification of the falsifier cluster in each iteration.

Pseudocode 1 shows the pseudocode of the algorithm taking a constraint on the size of the falsifier cluster into account.

The algorithm is run for a defined setting of number of interviewers, share of falsifiers and extent of falsifications (1:). Then, a substantial number B of bootstrap replications is performed (2: to 8:). For each replication a new synthetic dataset with the defined properties is generated (3:) (for the details of the bootstrap procedure, see Section 3 below). Based on the dataset, the vector of indicator values $\tilde{\mathbf{i}}_k$ is calculated for every interviewer (4:). Next, two clusters are identified by means of the constrained version of the optimization heuristic described above, ensuring that the cluster labeled falsifier cluster does not contain more than the predefined number of interviewers (5:). The performance of this clustering is evaluated based on the MCC. After all bootstrap replications have been carried out, the overall performance can be summarized, For example, by the mean of the MCC for a given set-up (8:).

In order to evaluate this method, different values of the constraint on the size s of the falsifier cluster are combined with different simulation setups, that is, different shares of falsifiers and falsified questions. In reality, however, one would need a rough idea of the falsifiers' share in the underlying dataset to introduce a plausible constraint. Optimally, a preselection criterion like the total value of pairwise distances is used to compare different constraints in a pretest. The corresponding results could be used to define the most preferable constraint. The development of such a pretest is left for future research.

Before turning to the data and results, a special case of the above procedure is introduced. It consists in limiting the size of the falsifier cluster to one. In this case, the global optimization algorithm can be replaced by an exact enumeration procedure to find the one interviewer resulting in the optimum for the objective function if labeled as a falsifier. Obviously, this approach is well suited neither for obtaining an estimate of the extent of falsifications in the sample nor for the identification of a larger share of such

Pseudocode 1 Pseudocode of bootstrap procedure for cluster analysis.

1: Define parameter settings: number of bootstrap replications B; size restriction for group of potential falsifiers s, parameters $n_1, \ldots, n_5$
2: **for** $b = 1$ to B **do**
3: Create artificial dataset: n_1 honest interviewers with n_2 real questionnaires; n_3 falsifiers with n_4 fabricated and n_5 real answers in all of their questionnaires
4: Calculate indicators for all interviewers
5: Conduct clustering analysis imposing the group size restriction s for the group of falsifiers
6: Store performance of clustering procedure and indicators for given dataset
7: **end for**
8: Summarize statistical information of performance over all B datasets

falsifiers. However, given the low computational cost, it might be a sensible first step in quality control to identify this extreme interviewer and conduct a follow-up on his or her data. We will also report results on the quality of this simplified procedure for the detection of only one potential falsifier in Section 4. The evaluation of this procedure will not be based on MCC, but simply on the frequency over all bootstrap replications, with which a falsifier is actually found in this one element cluster.

3. Data

To assess the methods' performance, a substantial number of datasets from surveys would be required. To this end both interviews collected by honest interviewers and interviews collected by faking interviewers should be contained and identified a priori. Although anecdotal evidence suggests a substantial prevalence of deviant behavior in surveys, such datasets are rarely available (Bredl et al. 2013). Typically, identified falsifications are removed from the dataset prior to further analysis. Publications based on the cleaned data do not contain information about the falsifications as it might provide a negative signal on the data quality. Falsifications which have remained undetected are still present in datasets, but cannot be used for the evaluation of our methods either, as no known benchmark is available.

For the present study, we resort to the results of a large-scale experiment conducted at the University of Giessen in 2012 providing both real and fabricated data (Menold et al. 2013). 78 students of the University of Giessen were recruited as interviewers. In a first stage of the experiment, they each conducted about ten real interviews using a questionnaire comprising sociodemographic information and questions about study subjects and on attitudes. The respondents were recruited randomly by the interviewers among other students. The quality of these real interviews was verified by controlling the tape recordings of all interviews. In a second stage of the experiment, each student was asked to fabricate another ten interviews in the laboratory. As input for these falsifications, the students were provided with a short sociodemographic profile of one of the respondents from the real data who was not interviewed by themselves. Making use of this profile, the students were asked to generate data which should replicate a real interview as close as possible. An additional monetary incentive for generating high quality falsifications was provided which was distributed to those interviewers generating data that could not be assigned to the falsifier cluster by the method proposed in Bredl et al. (2012). Given this explicit incentive, the interviewers' knowledge from conducting real interviews first and their knowledge about the group from which the respondents have been recruited, the experimental setup promotes the generation of high-quality falsifications. Hence, these falsifications might be more difficult to detect compared to "quick and dirty" approaches, which might be more common in some real settings if interviewers are aware of missing or weak supervision. In fact, the quality of assessment of interviewers by the methods discussed in the previous section has been substantially higher in the few applications to real data (Bredl et al. 2012; Storfinger and Winker 2013). Therefore, we consider the data used in this article as a worst-case (difficult to detect) scenario, but will discuss potential limitations of the dataset in the concluding section.

In our set-up, for each respondent we obtain a real interview conducted by one interviewer and a fabricated one provided by a different interviewer. Thus, starting with

the fabricated interviews conducted by one interviewer, in each interview some questions can be replaced by the real answers provided by the respondent in a real interview. This way, it becomes possible to generate synthetic datasets which contain interviews composed of actual answers to some questions and the falsifications provided by the interviewer to the other ones (in contrast, in De Haas and Winker (2014) complete real and fabricated interviews have been used for one interviewer). The share of these falsifications can be controlled when generating the synthetic data. Obviously, for the group of nondeviant interviewers, only data from the interviews actually conducted are employed. As a consequence, it is possible to control both the share of falsifiers and the extent to which their fabricated interviews comprise real and false data.

We are interested in how well the methods described in the previous section perform depending on the share of falsifiers and the extent of falsifications within the fabricated interviews. Obviously, this performance will depend on the specific selection of data from our experiment and, as the consequence of a random choice, has to be considered stochastic. Thus it is not sufficient to consider single synthetic datasets, and instead the analysis has to be repeated a large number of times for different random selections to allow for systematic conclusions.

Given that hardly any appropriate data are available from real surveys and generating suitable data through thousands of separate experiments of the type described above is not feasible either, the well-known resampling method known as bootstrap (Efron 1979, 1982) is used to generate synthetic datasets with the defined properties. The procedure comprises the following steps for each bootstrap iteration, that is, for the generation of a single synthetic dataset to be used in the analysis: first, a predefined number of interviewers is chosen randomly (with replacement) from all interviewers. This group will represent the honest interviewers. This means, for each of these interviewers a fixed number of real questionnaires is selected (with replacement) from the actual interviews conducted during the experiment by the corresponding interviewer. Finally, based on the resampled questionnaires, the indicator values are obtained. Second, another group of interviewers is generated in the same way as for the honest interviewers by selecting randomly (with replacement) a predefined number of interviewers from all interviewers. This second group is meant to represent falsifiers. Therefore, for each interviewer in this group, the actual data are compiled in a modified way. A fixed share of fabricated questions is assumed for all questionnaires selected for this interviewer. These partial falsifications are obtained by starting with a randomly selected (with replacement) fabricated interview provided by the respective interviewer during the experiment. Then, a number of questions corresponding to one minus the share of fabricated questions is randomly selected within the questionnaire. The answer to these questions is replaced by the corresponding real data collected by another interviewer, which provided the profile for the falsification. Finally, the indicator values for the falsifiers are calculated. This procedure allows to generate a large number of synthetic datasets with well-defined properties regarding the number of interviewers, the share of falsifiers, the number of questionnaires per interviewer and the extent of partial falsifications for the falsifiers. This makes it possible to analyze the performance of the clustering method for different scenarios based on – in the present study – 1,000 different samples for each scenario.

4. Results

Given this article's main focus on semifalsifications and the potential gains in discriminatory power of imposing size constraints on the falsifier cluster, a smaller set of experiments compared to the design in De Haas and Winker (2014) is conducted. Thus we take into account the substantially higher computational burden due to the repeated application of the TA heuristic for different constraints on the size of the falsifier cluster. In order to preserve comparability, the set of experiments is chosen as a subset of the original design. The details of the design are summarized in Table 1.

For all experiments, a number of 150 interviewers – similar to the original setting of the experiment – is used. In previous analysis, it was found that reducing the total number of interviewers typically helps to improve the discriminatory power. Thus the results presented here might be considered as lower bounds for the performance to be expected when the number of interviewers is small. Furthermore, the actual share of falsifiers in the dataset might affect the discriminatory power. We consider two settings for the share of falsifiers, namely six percent as a low and 50% as a substantial value. Obviously, the case of 50% falsifications might be considered an extreme setting. Finally, to study the impact of only partial falsifications, both a setting with 50% fabricated data in each questionnaire and one with 100% fabrication as in the experimental data and also as analyzed by Storfinger and Winker (2013) is used for comparison. In Appendix B, we consider shares of falsifiers of two percent, ten percent and 20%, respectively, and provide results for further shares of falsified questions, in particular for 25%, 70%, 75%, 80%, 85%, 90%, and 95%. Given that the main qualitative findings support those obtained for the main design, we do not comment on these additional cases in the text.

For the new parameter "size of the falsifier cluster", several values are also considered. Given that this size is unknown in real applications, a value of two percent stands for a low expectation about the prevalence of faking that is below the actual shares considered in the bootstrap simulations (six percent and 50%). The algorithm should exhibit a low false alarm rate in this case. The value of six percent corresponds to the actual number of falsifiers in one setting and still is much lower than the actual number for the other. 25% is an assumption that is high for the low-share setting, which has to result in a high rate of false alarms, while it is still low for the high faker-share setting. The highest value of 50% corresponds to the high faker setting. In addition to these four values, the algorithm is also run without restriction of the size of the falsifier cluster for comparability with the results in Storfinger and Winker (2013) and De Haas and Winker (2014).

A final setting, which might turn out to be interesting for practical applications, when the focus is just on finding some fakers, but not necessarily many or all of them, is given by

Table 1. Simulation settings for bootstrap runs.

Dimension	Original sample	Values for bootstrap			
No. of interviewers	156				150
No. of questionnaires per interviewer	~10				20
Share of falsifiers	50%			6%	50%
Faking share	100%			50%	100%
Constraint on size of falsifier cluster	n.a.	2%	6%	25%	50%

a constraint of the falsifier cluster to size one. This setting does not require an explicit optimization, as the global optimum can be easily identified by the full enumeration of all possible cases (i.e., by checking one interviewer at a time to represent the falsifier cluster, and selecting the one resulting in the best value of the objective function). It could be extended recursively by applying the method again after removing the interviewer identified as a potential falsifier in the previous step. Alternatively, one might look at the aggregated indicator values per interviewer and label the interviewer(s) exhibiting the largest values as potential falsifiers. For this approach, it is required to know a priori the direction of deviations of indicator values in falsifications, which might not always be obvious. Both approaches are not listed in Table 1, but we will also report results for these screening tools.

The results for the MCC values for all settings except the last mentioned screening devices are summarized in Figure 1. The bars labeled "w/o" provide the results for the unconstrained clustering, that is, using the same methodology as De Haas and Winker (2014). For the case of complete falsifications (graphs in the right column labeled "100% falsified questions") the results are comparable to the same setting in De Haas and Winker (2014). They confirm that the method has a strong potential to identify falsifiers, in particular if the share of falsifiers is high (lower right graph). If only few fakers are present (upper right graph), the tendency of the unconstrained clustering procedure towards clusters of similar sizes, results in a large share of false alarms and, consequently, in an MCC value close to zero.

The performance of the unconstrained clustering method deteriorates when only partial falsifications are considered (left column of Figure 1). While the performance remains weak for the case with few fakers (upper left graph), it becomes substantially worse for the high faker setting (lower left graph). In fact, the MCC value shrinks from about 0.5 to only 0.1. Again, this finding is qualitatively similar to those obtained by De Haas and Winker (2014) for the case of partial falsifications in the sense that a faker produces some completely real and some completely fabricated interviews.

Against this background, it is of interest to see to what extent the restricted cluster algorithm can improve the discrimination between honest and faking interviewers. When

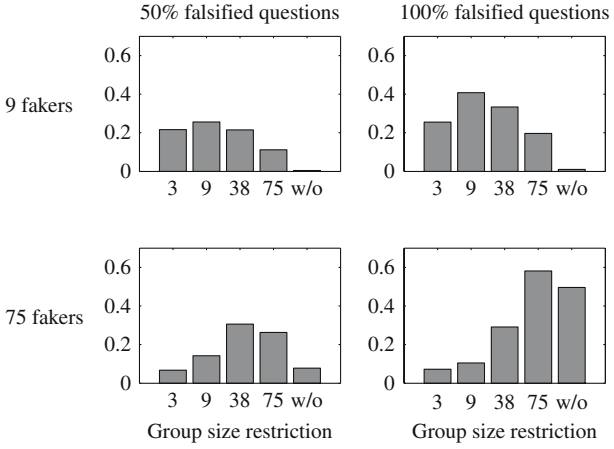

Fig. 1. Mean of MCC values over all bootstrap replications for different settings.

the correct restriction is imposed, that is, a number equal to the actual number of falsifiers (nine for the upper and 75 for the lower row), the MCC values always take on their maximum value, exhibiting a substantial improvement compared to the unconstrained version. These improvements are most pronounced for a low number of falsifications and for the case with partial falsifications. As long as the actual number of falsifiers is low, even wrong assumptions about this number still result in substantially improved MCC values both for partial and complete falsifications (upper row). In the second setting with a large number of falsifiers, however, imposing overly restrictive values for the size of the falsifier cluster (i.e., three or nine instead of the actual number of 75) results in a performance worse than the one of the unconstrained algorithm at least for completely fabricated questionnaires. In the case of partial falsifications, even the restriction to only nine falsifiers results in a slight improvement of the MCC value as compared to the unconstrained setting.

More insights into these results can be obtained by looking separately at the frequencies of oversights and false alarms, which are reported for the same settings in Figure 2. As expected, imposing a small size for the cluster containing the potential falsifiers results in a substantial share of oversights, in particular if the actual number of fakers is high. However, at the same time, the frequency of false alarms is remarkably low, suggesting that the method might be helpful to identify at least a subset of all falsifiers with some precision as long as the imposed size constraints are close to or smaller than the actual number of falsifiers.

We finish with a look at the "screening tools". Given that in all settings described in Table 1 at least one (partial) falsifier is present in the sample, when imposing a falsifier cluster of size one, one would hope that the method always spots one of the actual falsifiers. In fact, the simple procedure comes close to this result for the first interviewer marked as potential falsifier when considering only the more challenging setting of partial falsifications. If only nine falsifiers are present, in 81.6% of the bootstrap samples an

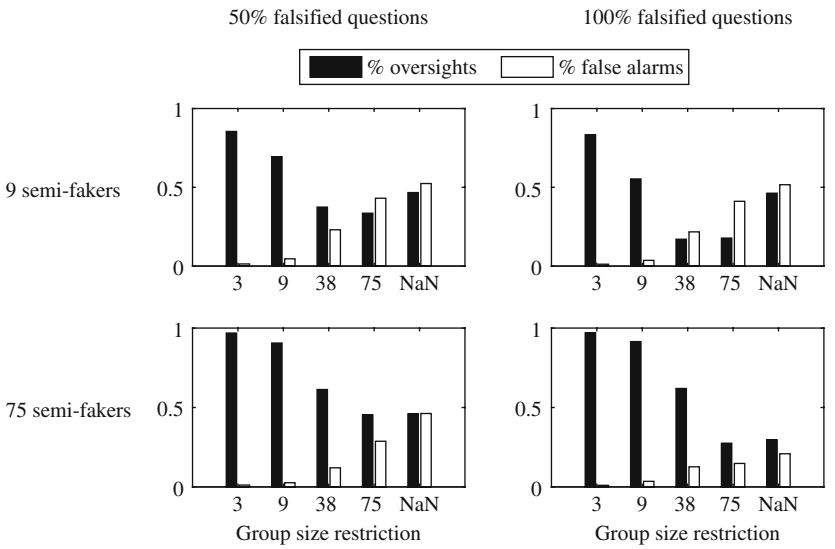

Fig. 2. Frequencies of oversights and false alarms over all bootstrap replications for different settings.

actual falsifier is found, while this share is 98.1% in the case of 75 falsifiers. The alternative approach of selecting the interviewers with highest aggregate indicator values results in shares of 65.5% and 99.9%, respectively. Obviously, if the sample does not contain a single falsifier, the one signal generated by the screening tools will always be a false alarm. Thus, it should not be used to "identify" falsifiers, but rather to select one case for a careful follow up if available resources do not allow for a more comprehensive quality check.

5. Conclusions and Outlook

The present article has analyzed how the identification of falsifiers by means of a data-driven clustering procedure is negatively affected if only partial falsifications are presented in the sense that falsifiers use partially real information and complement this with fabricated data. As expected, the performance deteriorates substantially when the share of fabricated answers decreases. This happens to a similar extent as for the setting when falsifiers generate some of their questionnaires completely (De Haas and Winker 2014).

The situation that interviewers rather fabricate some of their assignments or parts of the questionnaire than delivering only complete falsifications is considered as quite typical in real survey settings. Although empirical evidence is limited, completing partial interviews after a break-off by the respondent or just collecting basic sociodemographic information from the respondent and fabricating lengthy or sensitive parts of the questionnaire might represent situations resulting in partial fabrications. Therefore, to deal with the shortcomings of the previously proposed clustering method in this situation, a new clustering procedure is proposed, which allows imposing an a priori restriction on the number of falsifiers in the corresponding cluster. It is shown that imposing such restrictions improves the performance substantially. This holds in particular if the share of falsifiers is low, only partial falsifications are present, and the assumed share of falsifiers is close to the actual number.

As an extreme setting of this restricted clustering approach, a falsifier cluster of size one is also considered. While it is obvious that this method cannot produce a good overall assignment in a case where several falsifiers are present, it appears to be a valuable screening tool, as in all settings for most individual bootstrap replications a falsifier was correctly identified.

Given the high cost of classical methods of quality management such as reinterviews (Forsman and Schreiner 1991), it is recommended to apply the method presented here to select a small number of interviewers exhibiting conspicuous patterns. Thus the (restricted) size of the cluster containing the interviewers flagged for follow up might be set according to available resources for reinterviews or based on previous experience with prevalence of fabrications in a specific survey setting.

The present study has two major limitations, which might be overcome in future research. First, it might be argued that the experimental setting used to generate the data for the present analysis represents a worst-case setting in the sense that high-quality falsifications are obtained given the strong incentives for good fabrications in the experiment. At the same time, one might argue that the quality of falsifications was poor, given that all interviewers were students with no or limited previous experience as

interviewers. The prevailing effect might be evaluated making use of further experimental datasets. Second, falsifications in real data might differ from those obtained in an experimental setting. Therefore, further analysis based on real data such as in Bredl et al. (2012) is required. Besides dealing with these limitations, future research will also address some methodological issues. Alternative objective functions for the cluster construction will be considered, which might improve the performance in particular for the case of a low share of falsifiers, as the present method privileges clusters of about equal size. Furthermore, the usage of cross validation techniques to find a good a priori value for the size of the falsifier cluster is left for future analysis. Finally, probabilistic clustering methods are natural competitors in the case of partial falsifications and will also be a subject of our future work.

Appendix A

Indicators

Table 2 provides a summary of the indicators used to differentiate between data generated by interviewers following the prescribed procedure and data coming from faking interviewers. It provides the name of the indicator, a brief explanation of how it is constructed, the expectation about the sign of the deviation in its value between honest interviewers and falsifiers, a short argument explaining this expectation and a reference to the specific indicator. A more detailed description of all indicators used in the present study can be found in Menold et al. (2013) and De Haas and Winker (2014).

Table 2. Summary of indicators used to identify data potentially stemming from falsifications.

Indicator	Brief explanation	Higher value for	Rationale	Reference
Extreme Responding Style	Frequency of choosing extreme responses on rating scale	Honest	Falsifiers try to avoid extreme responses	Porras and English (2004)
Middle Responding Style	Frequency of choosing middle category (for uneven number of categories)	Falsifiers	Falsifiers try to avoid extreme responses	Storfinger and Opper (2011)
Acquiescent Responding Style	Agreement responses regardless of positive or negative item wording	Honest	Honest often agree regardless of content	Messick (1967)
Nondifferentiation	Standard deviation across all items	Falsifiers	Falsifiers use stereotypes	Reuband (1990)
News	Frequency of choosing fictitious categories (read magazines)	Falsifiers	Falsifiers try to save time and effort	Menold et al. (2013)
Filter	Frequency of choosing answers which allow skipping of further questions	Falsifiers	Falsifiers try to save time and effort	Hood and Bushery (1997)
Participation	Asking for past political activities	Honest	Falsifiers underestimate those activities	Menold et al. (2013)
Semi-Open	Frequency of choosing "other, please specify"	Honest	Falsifiers try to save time and effort	Hood and Bushery (1997)
Open	Frequency of providing answers to open questions	Honest	Falsifiers try to save time and effort	Bredl et al. (2012)
Rounding	Frequency of rounded numbers to numerical open questions	Falsifiers	Numerical information should be exactly known for real respondent	Tourangeau et al. (1997)
Primacy	Frequency of choosing the first two categories (visual presentation)	Honest	Honest tend to choose first option that seems satisfactory	Krosnick and Alwin (1987)
Recency	Frequency of choosing the last category (oral presentation)	Honest	Limited capacity of short-term memory	Krosnick and Alwin (1987)

Appendix B

Results for Other Shares of Falsified Questions

As additional information complementing Figure 1, we provide results for shares of falsified questions of 25%, 70%, 75%, 80%, 85%, 90%, and 95% in this appendix. Figure 3 shows the results for a situation when three falsifiers are present (two percent of all interviewers), while the results for 15 falsifiers (10%) and 30 falsifiers (20%) are exhibited in Figures 4 and 5, respectively.

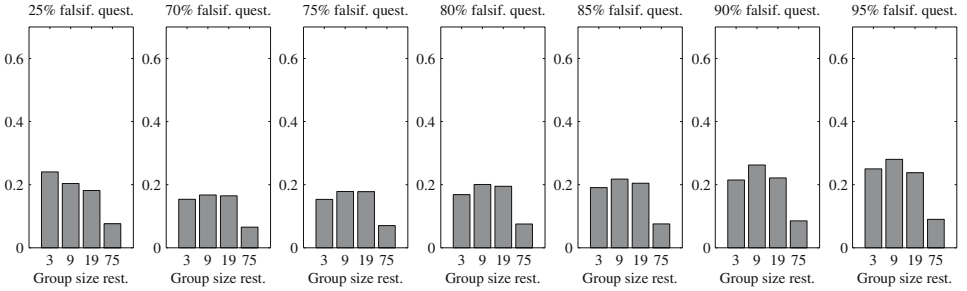

Fig. 3. Mean of MCC values over all bootstrap replications with three falsifiers.

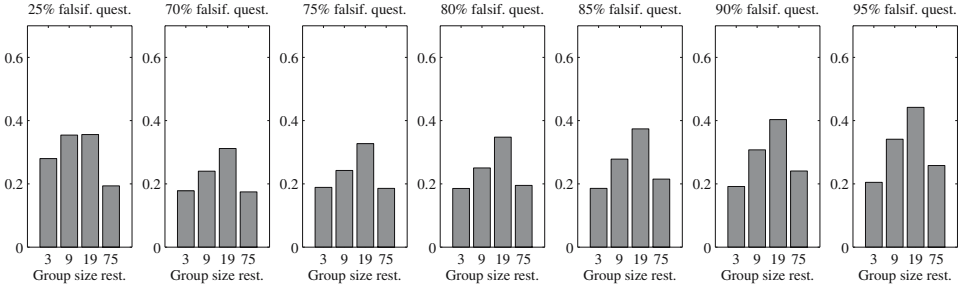

Fig. 4. Mean of MCC values over all bootstrap replications with 15 falsifiers.

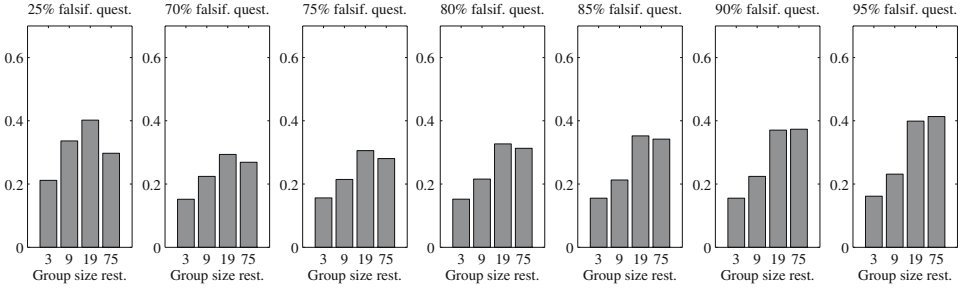

Fig. 5. Mean of MCC values over all bootstrap replications with 30 falsifiers.

6. References

Althöfer, I. and K.-U. Koschnik. 1991. "On the Convergence of Threshold Accepting." *Applied Mathematics and Optimization* 24: 183–195. Doi: http://dx.doi.org/10.1007/BF01447741.

Baragona, R., F. Battaglia, and I. Poli. 2011. *Evolutionary Statistical Procedures*. Statistics and Computing. Heidelberg: Springer.

Bredl, S., N. Storfinger, and N. Menold. 2013. "A Literature Review of Methods to Detect Fabricated Survey Data." In *Interviewers' Deviations in Surveys - Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 3–24. Frankfurt am Main: Peter Lang.

Bredl, S., P. Winker, and K. Kötschau. 2012. "A Statistical Approach to Detect Interviewer Falsification of Survey Data." *Survey Methodology* 38: 1–10.

Crespi, L. 1945. "The Cheater Problem in Polling." *The Public Opinion Quarterly* 9: 431–445.

De Haas, S. and P. Winker. 2014. "Identification of Partial Falsifications in Survey Data." *Statistical Journal of the IAOS* 30: 271–281. Doi: http://dx.doi.org/10.3233/SJI-140834.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7: 1–26. Doi: http://dx.doi.org/10.1214/aos/1176344552.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 38. Doi: http://dx.doi.org/10.1137/1.9781611970319.

Finn, A. and V. Ranchhod. 2013. "Genuine Fakes: The Prevalence and Implications of Fieldworker Fraud in a Large South African Survey." SALDRU Working Papers 115, Southern Africa Labour and Development Research Unit, University of Cape Town. Available at: http://ideas.repec.org/p/ldr/wpaper/115.html (accessed October 22, 2015).

Forsman, G. and I. Schreiner. 1991. "The Design and Analysis of Reinterview: An Overview." In *Measurement Errors in Surveys*, edited by P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman, 279–301. Chichester: Wiley. Doi: http://dx.doi.org/10.1002/9781118150382.ch15.

Gilli, M., D. Maringer, and E. Schumann. 2011. *Numerical Methods and Optimization in Finance*. Waltham, MA: Academic Press.

Gwartney, P. 2013. "Mischief Versus Mistakes: Motivating Interviewers to not Deviate." In *Interviewers' Deviations in Surveys - Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 195–215. Frankfurt am Main: Peter Lang.

Hood, C. and M. Bushery. 1997. "Getting More Bang from the Reinterviewer Buck: Identifying 'at Risk' Interviewers." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, August 10th to 14th 1997, Anaheim, CA, 820–824. Available at: https://www.amstat.org/sections/srms/Proceedings/papers/1997_141.pdf (accessed October 22, 2015).

Kemper, C. and N. Menold. 2014. "Nuisance or Remedy? The Utility of Stylistic Responding as an Indicator of Data Fabrication in Surveys." *Methodology: European*

Journal of Research Methods for the Behavioral and Social Sciences 10: 92–99. Doi: http://dx.doi.org/10.1027/1614-2241/a000078.

Kemper, C., V. Trofimow, B. Rammstedt, and N. Menold. 2011. "Indicators for the ex post Detection of Faking in Survey Data Constructed from Responses to the Big Five Inventory-10 (BFI-10)." Poster presented at the 11th European Conference on Psychological Assessment, date of conference, Riga, Latvia. Available at: http://www.ecpa11.lu.lv/files/Kemper Christoph.pdf (accessed October 22, 2015).

Krosnick, J. and D. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement." *Public Opinion Quarterly* 51: 201–219. Doi: http://dx.doi.org/10.1086/269029.

Matthews, B. 1975. "Comparison of the Predicted and Observed Secondary Structure of t4 Phage Lysozyme." *Biochimica et Biophysica Acta* 405: 442–451. Doi: http://dx.doi.org/10.1016/0005-2795(7590109-9).

Menold, N. and C. Kemper. 2014. "How Do Real and Falsified Data Differ? Psychology of Survey Response as a Source of Falsification Indicators in Face-to-Face Surveys." *International Journal of Public Opinion Research* 26: 41–65. Doi: http://dx.doi.org/10.1093/ijpor/edt017.

Menold, N., P. Winker, N. Storfinger, and C. Kemper. 2013. "A Method for ex-post Identification of Falsifications in Survey Data." In *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 25–47. Frankfurt am Main: Peter Lang.

Messick, S. 1967. "The Psychology of Acquiescence, an Interpretation of Research Evidence." In *Response Set in Personality Assessment*, edited by I. Berg. Chicago: Aldine Publishing Company. Doi: http://dx.doi.org/10.1002/j.2333-8504.1966.tb00357.x.

Porras, J. and N. English. 2004. "Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys." In *Proceedings of the Survey Research Methods Section: American Statistical Association*, August 8th to 12th 2004, Toronto, 4223–4228. Available at: http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000879.pdf (accessed October 23, 2015).

Reuband, K.-H. 1990. "Interviews, die keine sind, 'Erfolge' und 'Mißerfolge' beim Fälschen von Interviews." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 42: 706–733.

Schäfer, C., J. Schräpler, K. Müller, and G. Wagner. 2005. "Automatic Identification of Faked and Fraudulent Interviews in the German SOEP." *Schmollers Jahrbuch* 125: 183–193.

Storfinger, N. and M. Opper. 2011. "Datenbasierte Indikatoren für potentiell abweichendes Interviewerverhalten." Discussion Paper 58, ZEU, September 2011, Giessen. Available at: http://geb.uni-giessen.de/geb/volltexte/2012/8559/pdf/ZeuDiscPap58.pdf (accessed October 23, 2015).

Storfinger, N. and P. Winker. 2013. "Assessing the Performance of Clustering Methods in Falsification Using Bootstrap." In *Interviewers' Deviations in Surveys - Impact, Reasons, Detection and Prevention*, edited by P. Winker, N. Menold, and R. Porst, 49–65. Frankfurt am Main: Peter Lang.

Tourangeau, R., K. Rasinski, J. Jobe, B. Jared, T. Smith, and W. Pratt. 1997. "Sources of Error in a Survey on Sexual Behavior." *Journal of Official Statistics* 13: 341–365.

Verbiest, N., K. Vermeulen, and A. Teresdai. 2015. "Evaluation of Classification Methods." In *Data Classification – Algorithms and Applications*, edited by C. Aggarwal, 633–655. Boca Raton, FL: CRC Press.

Winker, P. 2001. *Optimization Heuristics in Econometrics: Applications of Threshold Accepting*. Chichester: Wiley.

Received April 2015
Revised October 2015
Accepted November 2015