Schmollers Jahrbuch 125 (2005), 183 – 193
Duncker & Humblot, Berlin

Automatic Identification of Faked and Fraudulent Interviews in the German SOEP

By Christin Schäfer, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G. Wagner

Abstract

Based on data from the German Socio-Economic Panel (SOEP), this paper presents two new tools for the identification of faked interviews in surveys. One method is based on Benford's Law, and the other exploits the empirical observation that fakers most often produce answers with less variability than could be expected from the whole survey. We focus on fabricated data, which was taken out of the survey before the data was disseminated to external users. For two samples, the resulting rankings of the interviewers with respect to their cheating behavior are given. For both methods all of the evident fakers are identified.

JEL Classifications: C8, C4

1. Introduction

1.1 Faking

In any survey in which the data are collected by personal interviews there is a danger that these interviewers will cheat. We can distinguish several forms of cheating: First, the most blatant form is when an interviewer fabricates all "responses" for an entire questionnaire. The U.S. Census Bureau refers to this practice as "falsification" or "fabrication". Falsification also includes the acceptance of proxy information when self-response is required and the unauthorized use of the telephone when a personal visit is required. A second, more subtle form of cheating is when an interviewer asks some questions in an interview and fabricates the responses to others. A third form of cheating is when an interviewer knowingly deviates from prescribed interviewing procedures, for example by conducting an interview with someone who is more easily reachable than the appropriate person and willing to participate in his or her place. In this paper we only address the first form of cheating, the fabrication of an entire interview.

1.2 Previous Findings on Cheating Behavior

Compared to other methodological topics, the literature contains only a few studies dealing with cheating by interviewers. Crespi (1945) investigated the factors that may contribute to cheating behavior. He distinguished between factors relating to questionnaire characteristics (design and length, difficult and antagonistic questions), administrative demoralizers (inadequate remuneration and training of the interviewer) as well as external factors (bad weather, bad neighborhoods, etc.). He proposed a twofold strategy of eliminating demoralizers. Furthermore he used a verification method to deter cheating. Some more recent studies refer to these verification methods and deal with optimal designs of quality control samples to detect interviewer cheating (Biemer and Stokes 1989) and the evaluation of quality control procedures for interviewers (Stokes and Jones 1989).

Because of the lack of factual information concerning the nature of interviewer falsification, the U.S. Census Bureau implemented an "Interviewer Falsification Study" in the year 1982 (Schreiner, Pennie, and Newbrough 1988). In this study, data was accumulated from fifteen surveys conducted by twelve U.S. Census Bureau regional offices over a five-year period. They found 205 cases of confirmed falsification. Most of these (74 percent) were detected through reinterviews and the majority (79 percent) was determined to have fabricated interviews. Their results provide evidence that the shorter the length of service, the more likely it is that an interviewer will falsify data (Schreiner, Pennie, and Newbrough 1988). Furthermore, when new interviewers falsify data, it is usually a relatively high proportion of their assignments and they tend to fabricate entire interviews. Interviewers with five or more years of experience usually falsify a smaller proportion of their assignments and tend to classify eligible units as ineligible (Hood and Bushery 1997).

Other studies like Reuband (1990), Schnell (1991) and Diekmann (2002) deal with the "quality" of faked interviews and the impact of fabricated data on substantive analysis. For example, Schnell (1991) performed a study in which he substituted 220 real interviews of the German General Social Survey (ALLBUS 1988, N = 3052) with fictive interviews and analyzed their effect on substantive results.

1.3 Fabrication within the German Socio-Economic Panel (SOEP)

In contrast to cross-sectional surveys, complex long-term panel studies like the SOEP (German Socio-Economic Panel Study) make it extremely difficult for interviewers to falsify data because the respondent is interviewed face to

face every year, and because a consistency check between waves reveals irregularities immediately. Hence we can assume that fabricated data will be a problem mainly in the first wave and will be detected quickly after conducting the second wave. In our case, the fieldwork organization gives us the faked records, whereas other fieldwork organizations hide this problem. They also provide hints about the standard quality control procedures performed in order to detect fakes. These verification methods as well as "conventional" statistical tests of stability and consistence are the ones proposed by Crespi (1945).

The SOEP consists of several samples with starting years ranging from 1984 though 2002 (Schupp and Wagner 2002). Fabricated data are rare and they have always been found in the first wave of each sample (with the exception of the East German sample C and the small sample D, which are "clean"). Only one interviewer was able to fabricate data for the first two waves without raising suspicion until wave 3 (Sample E). The first wave of samples A and B contains only 0.6 and 1.5 percent fabricated data, respectively, and the first wave of sample E contains about 2 percent faked household interviews. In the second wave, approximately 1 percent of fabricated data was identified in sample E. In the first wave of sample F, only 0.1 percent of the interviews were detected as fabricated. This share equals 11 records. Due to this small number of cases, only samples A / B and E will be analyzed.

Because Biemer and Stokes (1989) find that in two large demographic surveys, cheating behavior differed between urban and rural areas, we examine these kind of differences. The results are not consistent: for sample A / B the area effect is significant on a 1 percent level ($\chi^2 = 1452$), whereas in sample E the existence of an area effect can not be shown ($\chi^2 = 0.06$).

Only very little is known about the characteristics of interviewers who cheat in surveys. Koch (1995) shows that younger interviewers with a higher educational level have more inconsistencies in their interviews than others. All interviewers who fabricated data (N = 9) in the SOEP are middle-aged males. We find no education effects. In addition in sample A, cheating interviewers have on average a higher assignment of household interviews (18.3) than the interviewers in the non-faked data (9.6). In sample E, the difference between the average assignments (non-faked data: 7.32; faked data: 11.67) is not statistically significant. In the first wave of all samples, almost all cheating interviewers falsified their entire assignments, and only one interviewer in samples A and B falsified just one interview, out of a total of 43 personal interviews. All of these interviewers were working on this panel study for the first time. We can assume that they were not aware of the effectiveness of quality control in SOEP, or of the fact that fakes in the panel design are easily identifiable through consistency checks over two waves. Because these checks cannot be conducted on cross-sectional surveys, we seek methods that can identify fabricated data with a "one-shot procedure".

2. Two New Methods for Fraud Detection in Surveys

2.1 Benford's Law

Benford's Law is an empirical "law" which states that in many tables of numerical data, the leading digits are not uniformly distributed as might be expected, but rather obey a certain logarithmic probability distribution. Benford (1938) derived a formula to predict the frequency of numbers found in many categories of tables. The leading (non-zero) digit obeys the law

$$(1) \qquad Prob\ (\text{first significant digit} = d) = log_{10}\left(1 + \frac{1}{d}\right),$$

for $d = 1, 2, \ldots, 9$. Hence, in a number chosen at random, the leading digit $d = 1$ tends to occur with probability 0.301, leading digit $d = 2$ with probability 0.176, and so on monotonically down to probability 0.046 for leading digit $d = 9$. For many years, the status of this law was little more than a numerical curiosity, but practical implications began to emerge in the 1960s (Scott / Fasli 2001).

A plausible theoretical explanation for the appearance of this logarithmic distribution is the random-samples-from-random-distribution theorem by Hill (1995). He shows that "if probability distributions are selected at random, and random samples are then taken from each of these distributions in any way so that the overall process is scale (or base) neutral, then the significant digit frequency of the combined sample will converge to the logarithmic distribution." (Hill 1995, p. 360). It is not required that individual realizations of a random variable be scale- or base-invariant. But it is necessary that the sampling process on average does not favor one scale over another.

This theorem gives the answer to the question whether Benford's Law can be applied to survey data, because survey data contain different variables with different distributions. Therefore we can test whether the chosen mixture of variables from survey data are scale-unbiased. If this is the case, it is reasonable that this mixture of data follows Benford's Law.

2.2 Results with Benford's Law

First we provide a description of the data we examine using Benford's Law. The selected data are restricted to variables with monetary values. Besides monthly gross and net labor income, the data sets contain variables like the gross amount of Christmas or vacation bonus, gross amount of monthly unemployment benefits or monthly subsistence allowance, gross amount of early

retirement benefits, amount of taxes, as well as many other monetary variables.

The estimated leading digit distributions for the first wave of sample A / B and the first two waves of sample E have almost the same shape. The distributions are unimodal and the medians are always lower than the means, leading to positive skewed distributions. A unimodal positive skewed distribution is one important requirement for the use of Benford's Law (Scott / Fasli 2001).

We have shown that those interviewers who do falsify data in fact fabricate a large proportion of their assignments. Therefore in order to increase the statistical power of our analysis, we analyze whole clusters of interviews per interviewer ("interviewer cluster") rather than individual questionnaires. If real survey data follows the logarithmic distribution and fabricated survey data does not, we should be able to identify these clusters of fabricated interviews and to test them for significance.

To explore the fit of each cluster we calculate χ^2 values

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{(h_{d_i} - h_{b_d})^2}{h_{b_d}} ,$$

where n_i is the number of first digits in the interviewer cluster i, h_{d_i} is the observed proportion of digit $d = 1, \ldots, 9$ in interviewer cluster i and h_{b_d} is the proportion of digit d under Benford's distribution. Since the χ^2 values depend on the number of observations, we calculate the probability for the realized χ^2 values with a bootstrap method.

An approximation of the probability of obtaining a value of the χ^2-statistic more extreme than that actually observed, $Prob(\theta > \hat{\theta})$, can be obtained directly from the proportion of bootstrap replications B higher than the original estimate $\hat{\theta}$. These probabilities reflect the plausibility of the fit to Benford independent of the number of digits in the cluster. Our hypothesis is that cheating interviewers have very low probabilities. Hence we construct an interviewer-ranking by probability values.

Table 1 shows the top of the ranking list for the first wave of samples A / B (636 interviewers) and E (150 interviewers). The known faking interviewers are marked. We see that several cheating interviewers occur on the top of the list because their fit statistics are not plausible. If we look at the first ten interviewers as suspicious, with Benford we identify one out of three fakers in sample A, and in sample E, three out of five fakers.

Table 1

Interviewer ranking with Benford (faking interviewers marked)

Sample A / B, wave 1, n = 636}					Sample E, wave 1, n = 150				
Rank	Int.no	digits	χ^2	plausibility	Rank	Int.no	digits	χ^2	plausibility
1	**xx279x**	**122**	**52.30**	**0.0020**	**1**	**xx837x**	**221**	**49.07**	**0.0030**
2	xx147x	94	46.88	0.0040	2	xx328x	61	42.58	0.0140
3	xx856x	28	28.48	0.0060	3	xx665x	40	40.08	0.0170
4	xx500x	32	23.95	0.0180	**4**	**xx289x**	**158**	**52.16**	**0.0260**
5	xx878x	29	21.56	0.0410	5	xx796x	177	43.48	0.0430
6	xx320x	16	28.01	0.0450	6	xx908x	27	32.22	0.0930
7	xx003x	45	25.50	0.0470	**7**	**xx281x**	**7**	**28.15**	**0.1030**
8	xx363x	46	25.37	0.0510	8	xx674x	85	30.14	0.1440
9	xx097x	25	22.51	0.0630	9	xx690x	173	35.62	0.1750
10	xx687x	27	19.34	0.0680	10	xx059x	18	23.60	0.1630
11	xx425x	94	26.19	0.0800	11	xx085x	136	37.32	0.1940
12	xx830x	20	21.22	0.0890	12	xx613x	143	34.36	0.2050
13	xx563x	33	19.18	0.0930	13	xx184x	71	30.04	0.2170
14	xx566x	58	31.81	0.0970	14	xx370x	271	33.66	0.2080
15	xx016x	26	19.35	0.1000	15	xx901x	137	34.49	0.2360
16	xx353x	4	18.24	0.1000	16	xx899x	109	31.60	0.2790
17	xx525x	24	20.69	0.1020	17	xx376x	89	25.23	0.2720
18	xx208x	33	18.62	0.1040	18	xx335x	41	22.28	0.2860
19	xx654x	226	41.93	0.1040	19	xx937x	9	23.92	0.3280
20	xx632x	36	19.09	0.1080	20	xx608x	258	25.33	0.3080
21	xx846x	33	18.43	0.1090	21	xx424x	13	19.27	0.3570
22	xx877x	33	18.09	0.1190	22	xx441x	83	24.78	0.4490
23	xx841x	11	23.76	0.1200	23	xx761x	178	25.93	0.4720
24	xx085x	37	20.14	0.1220	24	xx534x	105	26.91	0.4740
25	xx760x	170	42.35	0.1260	25	xx818x	90	21.15	0.5020
26	xx365x	45	21.13	0.1340	26	xx689x	81	25.95	0.4850
27	xx066x	7	22.00	0.1380	27	xx118x	159	26.91	0.5360
28	xx200x	37	19.50	0.1430	28	xx907x	103	26.05	0.5280
29	xx650x	29	17.15	0.1440	29	xx393x	84	22.87	0.4970
30	xx052x	24	18.81	0.1540	30	xx785x	111	24.20	0.5340
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Int.no.: number of interviewers, digits: number of digits in cluster.

Source: SOEP, individual questionnaire, only monetary variables (own calculation).

2.3 The Variability Method

The variability method is based on the empirical evidence that the variance of all answers across all questionnaires delivered by a faking interviewer is lower than the variance achieved by questionnaires of non-fabricated interviews. There are several points that could explain the reduction in, or even the complete absence of variance in fabricated interviews:

- Fakers tend to answer every question. Thus they produce less missing values.
- In questions where one needs to assign a score, for example from (1) "I agree" to (5) "I disagree", fakers tend to make a check mark in the middle. Extreme values are avoided.
- Since the interviewers know the questionnaire and understand the meaning of the questions, they will not produce any astonishing answers when faking. Such answers can be found in non-fabricated interviews because the interviewees have misunderstood a question.

The variability method consists of the following steps: first, measure the variance within all the questionnaires of one interviewer; second, compare this value to the expected variance for a questionnaire cluster of the given size on the whole survey. More formally, let I_i, $i = 1, \ldots, n$, denote the interviewer i, and n is the number of interviewers that have conducted the survey. The number of questionnaires Q_j is given by m with $j = 1, \ldots, m$ and $m = m_1 + \ldots + m_i$, where m_i denotes the number of questionnaires delivered by interviewer I_i. Without taking into account any meaning of the answers – whether a 5 stands for "5 years" or for "I disagree" – we calculate the variance for every question $Q(k)$, $k = 1, \ldots, l$ on all questionnaires Q_j of an interviewer I_i and total results across all questions:

$$T_{I_i} = \sum_{k=1}^{l} \sum_{j=1}^{m_i} (Q_j(k) - \overline{Q(k)})^2 \; . \tag{3}$$

Here, $\overline{Q(k)}$ denotes the mean for question $Q(k)$ and the index j accounts for all questionnaires $Q_j, j = m_{i1}, \ldots, m_{im_i}$ of interviewer I_i.

The distribution of the test statistic T is estimated using a resampling approach on the whole survey. From this distribution, we can derive a probability of the observed value. In the following we will denote this probability with plausibility. By sorting the interviewers with respect to the plausibility they achieved, we obtain an interviewer ranking. The interviewers with the lowest plausibility are at the top of the ranking. They are considered to be potential fakers.

The procedure is defined as follows: The value of T_i (as defined in equation 2), which is assigned to interviewer I_i, is compared to the corresponding distribution of the test statistic T, which is estimated using a resampling approach. The area under the density curve on the left side of the realization T_i defines the plausibility. If the plausibility is too small, the interviewer is considered to be a potential faker. The procedure corresponds to a one-sided statistical test. One could argue that interviewers who achieve a plausibility that is suspiciously large could be fakers as well. Following this argument, one has to conduct a two-sided test. However, from the results of our experiments we conclude that this argument does not hold and that for the given task, a one-sided statistical test is more appropriate.

2.4 Results with the Variability Method

In Table 2 the interviewer rankings for sample A / B and sample E, wave 1 are shown. Interviewers who achieve the same plausibility value are sorted in increasing order of their personal identification number. The known fakers appear on top of the rankings. It is remarkable that interviewer xx289x, who had faked questionnaires in two waves of sample E and who was detected only in the third wave, is immediately debunked with the variability method as well as with the Benford method in wave 1. Notice that for sample A / B, the variability method is a little bit more effective than the Benford test.

3. Discussion

The data basis consists of raw data from the German Socio-Economic Panel (SOEP). A total of 90 faked household interviews and 184 faked individual interviews were detected by conventional verification methods such as reinterviewing, almost all of them after the first wave of a subsample. The share of fabricated data is low in all samples (far less than 1 percent) and the maximum is 2.4 percent in sample E. In subsamples C and D, no fakes were identified. It is important to note that except for the fakes in sample E, faked data were never disseminated within the widely-used SOEP, because fakes were detected and deleted from the database prior to its release to external users. But those fakes that were contained in the original data files provided by the fieldwork organization are kept at DIW Berlin and provide a rich source for methodological research.

We applied two new approaches for discovering frauds which do not require two waves of data but can be applied to cross-sectional data as well. First we utilized a procedure based on Benford's Law to survey data and used it for fraud detection in the SOEP. Second we developed a new method we call the

Table 2

Interviewer ranking with the variability method (faking interviewers marked)

Sample A / B, wave 1, n = 636				Sample E, wave 1, n = 150			
Rank	Int.no.	Q.no.	plausibility	Rank	Int.no.	Q.no.	plausibility
1	xx306x	25	0.00254	1	xx202x	25	0.00000
1	xx111x	222	0.00254	1	xx118x	27	0.00000
1	xx766x	40	0.00254	1	xx901x	29	0.00000
1	xx012x	89	0.00254	**1**	**xx289x**	25	0.00000
1	xx856x	18	0.00254	1	xx690x	29	0.00000
1	xx441x	29	0.00254	1	xx665x	12	0.00000
7	xx870x	22	0.00252	7	xx201x	10	0.00024
8	**xx378x**	32	0.00254	**8**	**xx281x**	2	0.00044
9	xx343x	35	0.00258	9	xx441x	19	0.00054
10	**xx279x**	35	0.00259	9	xx240x	24	0.00054
11	xx370x	119	0.00266	11	xx820x	25	0.00064
12	xx145x	22	0.00281	12	xx290x	71	0.00114
13	xx624x	64	0.00317	13	xx273x	22	0.00164
14	**xx800x**	38	0.00323	13	xx502x	18	0.00164
15	xx916x	13	0.00338	15	xx907x	27	0.00174
16	xx320x	6	0.00344	16	xx328x	15	0.00224
17	xx440x	13	0.00345	**17**	**xx837x**	32	0.00324
18	xx901x	14	0.00363	18	xx370x	49	0.00384
19	xx161x	11	0.00373	19	xx086x	2	0.00634
20	xx104x	66	0.00382	20	xx275x	15	0.00714
21	xx704x	2	0.00399	21	xx145x	14	0.00874
22	xx460x	8	0.00427	22	xx862x	27	0.00914
23	xx473x	33	0.00443	23	xx376x	12	0.01074
24	xx187x	6	0.00445	24	xx393x	15	0.01174
25	xx747x	60	0.00474	25	xx921x	9	0.01344
26	xx206x	24	0.00477	26	xx160x	13	0.01674
27	xx093x	11	0.00506	27	xx904x	3	0.02554
28	xx730x	30	0.00549	28	xx691x	8	0.02724
29	xx340x	10	0.00579	29	xx689x	19	0.02774
30	xx160x	29	0.00599	30	xx330x	9	0.03714
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Int.no.: number of interviewers, Q.no.: number of questionnaires.

Source: SOEP, individual questionnaire (own calculation).

variability method, which exploits the empirical observation that fakers most often produce answers with less variability than could be expected from the whole survey.

In both procedures, we derived test statistics for each interviewer cluster. The distributions of these test statistics were estimated using resampling approaches across the whole survey. From these distributions, we derived probabilities of the observed values. Then the interviewers were sorted with respect to the probabilities or plausibilities they achieved. From this, interviewer rankings were obtained. The interviewers with the lowest plausibility are at the top of the ranking. They are considered to be potential fakers.

We show that with both the Benford and the variability method, we can identify almost all of the clusters of fabricated interviews which we know to have been faked.

As logical next step, we explore the impact of faked and suspicious interviews on substantive research questions like the analysis of labor earnings. Due to space constraints, these findings are not reported here. The interested reader is referred to Schräpler / Wagner (2005) which describes some of the findings. Further information is available from the authors on request. In summary, we find empirical evidence for the finding of Schnell (1991) that even small proportions of faked interviews can be an important problem in multivariate survey statistics.

References

Benford, F. (1938): The Law of Anomalous Numbers. Proceedings of the American Philosophical Society, 78(4), 551 – 572.

Biemer, P. / *Stokes,* S. (1989): The Optimal Design Quality Control Samples to Detect Interviewer Cheating. Journal of Official Statistics, 5 (1), 23 – 39.

Crespi, L. (1945): The Cheater Problem in Polling. Public Opinion Quarterly, Winter, 431 – 445.

Diekmann, A. (2002): Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Manuskript 06 / 2002, Institut für Technikfolgenabschäatzung (ITA). Wien.

Hill, T. P. (1995): A Statistical Derivation of the Significant-Digit Law. Statistical Science 10, 354 – 362.

Hood, C. C. / *Bushery,* J. M. (1997): Getting more Bang from the Reinterview Buck: Identifying 'At Risk' Interviewers. Proceedings of the American Statistical Association (Survey Research Methods Section), 820 – 824.

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

Reuband, K.-H. (1990): Interviews, die keine sind – 'Erfolge' und 'Misserfolge' beim Fälschen von Interviews. Kölner Zeitschrift für Soziologie und Sozialpsychologie 4, 706 – 733.

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

Schräpler, J.-P. / *Wagner,* G. G. (2005): Characteristics and Impact of Faked Interviews in Surveys – An analysis of genuine fakes in the raw data of SOEP. Accepted for publication in: Allgemeines Statistisches Archiv, forthcoming.

Schreiner, I. / *Pennie,* K. / *Newbrough,* J. (1988): Interviewer falsification in Census Bureau Surveys. Proceedings of the American Statistical Association (Survey Research Methods Section), 491 – 496.

Schupp, J. / *Wagner,* G. G. (2002): Maintenance of and Innovation in Long-term Panel Studies The Case of the German Socio-Economic Panel (GSOEP) Allgemeines Statistisches Archiv 86(2), 163 – 175.

Scott, P. / *Fasli,* M. (2001): Benford's Law: An Empirical Investigation and a Novel Explanation. CSM Technical Report 349, Department of Computer Science, University Essex.

Stokes, L. S. / *Jones,* P. (1989): Evaluation of the Interviewer Quality Control Procedure for the Post-Enumeration Survey. Proceedings of the American Statistical Association (Survey Research Methods Section), 696 – 198.