Allgemeines Statistisches Archiv 89, 7–20
© Physica-Verlag 2005, ISSN 0002-6018

Abhandlungen / *Articles*

Characteristics and impact of faked interviews in surveys – An analysis of genuine fakes in the raw data of SOEP

By Joerg-Peter Schraepler and Gert G. Wagner[*]

Summary: Panel data offers a unique opportunity to identify data that interviewers clearly faked by comparing data waves. In the German Socio-Economic Panel (SOEP), only 0.5 percent of all records of raw data have been detected as faked. These fakes are used here to analyze the potential impact of fakes on survey results. Due to our central finding the faked records have no impact on the mean or the proportions. However, we show that there may be a serious bias in the estimation of correlations and regression coefficients. In all but one year (1998), the detected faked data have never been disseminated within the widely-used SOEP study. The fakes are removed prior to data release.

Keywords: Cheating, falsification, faked interviews, quality control, SOEP. JEL C8, C4.

1. Introduction

In any survey in which the data are collected by personal interviews there is a danger of cheating by interviewers, i.e. that some interviewers may fabricate data.[1] In principle we can distinguish three forms of cheating.

First, the most blatant form occurs when an interviewer fabricates all 'responses' for an entire questionnaire. The U.S. Bureau of the Census refers to this as 'falsification' or 'fabrication'. A more subtle form is the second one, 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 a different person who is more easily reachable and willing to take the place of the appropriate person. Falsification might also include the acceptance of proxy information when self-response is required and the unauthorized use of a telephone instead of a personal visit is required.

In our paper we deal only with the first form of cheating: The fabrication of an entire interview. We focus on faked data in the German Socio-Economic Panel (SOEP), examining both the potential and empirically

Received: 10.12.2003 / Revised: 09.08.2004

[*] We are grateful to participants in the workshop on 'Item Nonresponse and Data Quality on Large Social Surveys' for useful critique and comments, especially Rainer Schnell and our outstanding discussant Regina Riphahn. The usual disclaimer applies.

[1] Here we do not address cheating by respondents who do not tell the 'truth'.

demonstrable impacts of cheating on survey results. Since quality control mechanisms are not the focus of this study, we only very briefly describe the procedures used to detect the fakes.[2] In Section 2, we give an overview of other studies on cheating, and in Section 3 we describe the quality control procedures. Section 4 contains the main findings and Section 5 summarizes.

2. Previous findings on cheating behavior

Compared to other topics of statistical methodology, there are only very few studies dealing with cheating by interviewers in the literature. Crespi (1945) described several 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 dual strategy of eliminating demoralizers and using a verification method to deter from 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 the 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 *et al.*, 1988). In this study data were 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 %) were detected through reinterviews and the majority (79 %) was determined to have faked interviews.[3] Their results provide evidence that the shorter the length of service, the more likely an interviewer will falsify data (Schreiner *et al.*, 1988). Furthermore, when new interviewers fake data, it is usually a relatively high proportion of their assignments and they tend to fake 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 deal with the 'quality' of faked interviews and the impact of faked data on substantive analysis. Reuband (1990) shows that students are able to reproduce data in fictive interviews using given demographic variables of real respondents (in a study on Germany).

Using simulations, Schnell (1991) found only minor effects on multiple regressions in the case of less than 5 % fakes, but a serious impact on multivariate statistics and data quality given higher proportions. In this paper,

[2] However, these verification methods are the subject of few but important other studies dealing with fakes in surveys (cf. Biemer and Stokes, 1989; Stokes and Jones, 1989; Bushery *et al.*, 1999).

[3] The second most prevalent type of falsification (18.5 %) was deliberately misclassifying units as vacant when they were occupied (Schreiner *et al.*, 1988).

we intend to support these simulation-based findings by analyzing real faked data.

3. Detecting cheating interviewers in the SOEP

In contrast to cross-sectional surveys, successful falsification – from the interviewer's point of view – is extremely difficult in panel studies such as the German Socio-Economic Panel Study (SOEP) because the respondent is interviewed face to face every year, and because a consistency check between waves shows irregularities immediately. SOEP is a household panel survey that is used in analyses worldwide, as described by Wagner *et al.* (1993) and more recently by Schupp and Wagner (2002).[4]

SOEP was started in 1984 with two subsamples (A and B), which represented the German and the foreign population of West Germany. In 1990 after the fall of the Berlin wall, SOEP was extended to East Germany (subsample C). In 1994/95, a special sample of immigrants was introduced (subsample D), and in the years 1998 and 2000, the sample size was increased (subsamples E and F). Thus the different starting points and subsamples within SOEP make it possible to detect fakes because we know that most of them are identified after a first wave.

In a panel survey like SOEP we can assume that faked data will be a problem mainly in the first wave and will be detected quickly after conducting the second wave.[5] The results of quality control show that this was clearly the case in SOEP.

The most common method used for detecting interviewer cheating in face-to-face surveys is the 'verification method', where a sample of an interviewer's assignment is recontacted in order to verify that an interview was conducted (Biemer and Stokes, 1989; Schreiner *et al.*, 1989). Along this line, the German SOEP provides a unique opportunity to identify faked data. Falsifications are detected in several ways:

1. Most fakes can be identified easily by comparing the data from two waves. If one year's data deviates considerably from that of the previous year(s), the interview control department contacts the respective households by phone and asks the household members to verify the data.
2. After the interview, all respondents receive a 'thank-you' letter and a small gift by mail for having given the interview. Hence, if the interview did not take place, at least some households are likely to contact the

[4] The study is directed by the German Institute for Economic Research (DIW Berlin) and its fieldwork has been done for the last 20 years by a private fieldwork organization, Infratest Sozialforschung. The fieldwork procedures of Infratest Sozialforschung are certified by ISO 9000. The fieldwork organization uses mainly experienced interviewers for the SOEP project. The average length of service in the first wave is approximately five years.

[5] In SOEP the same housing unit is interviewed each year, and personal visits are required (or at least phone contact by the headquarters of the fieldwork organization). Hence, the risk of being caught falsifying data is relatively high.

fieldwork organization, which then becomes aware of the falsified interviews.

3. Due to problems with an above-average share of cheating interviewers in sample E of SOEP (1998), for sample F (2000), all households were recontacted after interviewing and asked to verify the household composition.

Once a fake is detected, it is erased from the analysis data files. This means that DIW Berlin received the faked data from wave 1 that was identified after a second wave took place. This faked data can be analyzed. In addition, one interviewer was able to fabricate data for the first two waves of sample E without raising suspicion until wave 3.

Table 1 shows the (detected) amount of faked data. The first wave of samples A and B contains only 0.6 and 1.5 % faked data, respectively, and the first wave of sample E contains about 2 % faked household interviews. In the following wave, approximately 1 % faked data was identified in sample E. In the first wave of sample F, cheating was minimal thanks to an improvement in fieldwork procedures: Only 0.1 % of the interviews were detected as faked. For subsamples C and D, no faked data was detected. With the exception of two years of subsample E faked data is deleted in data set, made available for analysis.

	Household interview			*Personal interview*		
Sample	Valid cases	Faked cases	Fakes in percent of total cases	Valid cases	Faked cases	Fakes in percent of total cases
1984						
Sample A	4 528	26	0.6	9 115	59	0.6
Sample B	1 393	22	1.5	3 175	45	1.4
1998						
Sample E	1 056	23	2.1	1 910	47	2.4
1999						
Sample E	886	11	1.2	1 629	22	1.3
2000						
Sample F1	5 848	8	0.1	10 470	11	0.1
Total (including samples C and D)	16 412	90	0.5	31 830	184	0.6

Source: SOEP 1984–2000.

Table 1. Proportion of detected faked raw data in the SOEP.

3.1. Area and interviewer characteristics for detected faked data in the SOEP. Biemer and Stokes (1989, p. 25) find that in the two large U.S. demographic surveys cheating behavior is more concentrated in urban areas. Table 2 shows the frequency of faked household interviews in

SOEP samples A, B, and E by number of residents in the area.[6] In the faked sample A/B, the distribution of the area characteristics differ from that in the non-faked sample. The proportion of falsification in cities (≥ 100 000 residents) is 52.1 % in sample A/B, and the proportion of cities in the non-faked data of sample A/B is only 40.4 %. Also the proportion of rural areas is higher in the faked sample than in the non-faked sample. These differences are statistically significant on a 1 % level (Chi-square = 1 452). Nevertheless, we find no statistically significant area effect in sample E if we only differentiate between cities and non-cities (Chi-square = 0.06). This finding suggests that if an area effect exists, it is not systematic. This may be important because the unknown true data faked by the cheating interviewers has the same area distribution as the faked interviews. Systematic differences between faked and non-faked interviews in the area characteristics would suggest that the distributions of the unknown true data are different from the known non-faked data.

Number of residents in area	A/B non-faked		A/B faked		E non-faked		E faked		Total non-faked		Total faked	
	N	%	N	%	N	%	N	%	N	%	N	%
≥ 100 000	2 369	40.4	25	52.1	375	35.5	12	35.3	2 744	39.6	37	45.1
20–100 000	1 543	26.3	1	2.1	240	22.7	22	64.7	1 783	25.7	23	28.0
≤ 20 000	1 959	33.4	22	45.8	441	41.8	-	0.0	2 400	34.7	22	26.9
Overall	5 871	100	48	100	1 056	100	34	100	6 927	100	82	100

Source: SOEP samples A, B, and E household questionnaire.

Table 2. Distribution of fakes in the first waves of sample A, B, and E by area characteristic.

Only very little is known about the characteristics of interviewers who cheat. Koch (1995, p. 97) shows that younger interviewers with a higher education level have more inconsistencies in their interviews than others. In SOEP, all interviewers who faked data ($N = 9$) are middle-aged males. We do not find education effects; cheating interviewers may have a university degree or only a primary school education. 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 at a 10 % level. In the first wave of all samples, almost all cheating interviewers falsified their entire assignments. Only one interviewer in samples A and B faked just one out of over 43 personal interviews. Whereas the first interviewers in sample A/B have been working for the fieldwork organization for more than three years, the latter one with only one faked interview has been there for only two years. All of them were working for this panel study for the first time. We can assume that they were not aware of the effectiveness of the quality control in this study or of the fact that its design makes fakes easy to identify. Our finding is not very different from the situation described by

[6] We don't give area characteristics for sample F because only one cheating interviewer could be detected.

Hood and Bushery (1997), who found that cheating interviewers who falsify a relatively high proportion of their assignments are inexperienced.

3.2. An indicator for faking: Stability between consecutive waves.

In SOEP, cheating is usually detected immediately after the first wave of a subsample. However, one interviewer was able to fabricate data for the first two waves of sample E. Hence, we are able to investigate whether the non-faked and faked answers are consistent from wave to wave or whether there is a difference between the non-faked and the faked data. Table 3 shows the stability coefficients (correlations) for several satisfaction items and items about worries. Furthermore, z-values are given to indicate whether the differences between the faked and non-faked correlations are statistically significant.[7] In the non-faked data set, the stabilities have positive values from 0.35 to 0.60; in the faked sample the stabilities are often close to zero, with only three values over 0.4 and two even negative. Half of the differences of the satisfaction stabilities and one-third of the worry stabilities are at least significant at a 5 % level. We can also see that the correlation in the total sample is in all cases very close to the correlations in the non-faked sample.[8]

4. Bias due to interviewer cheating

4.1. Analytical estimates of the possible bias due to interviewer cheating.

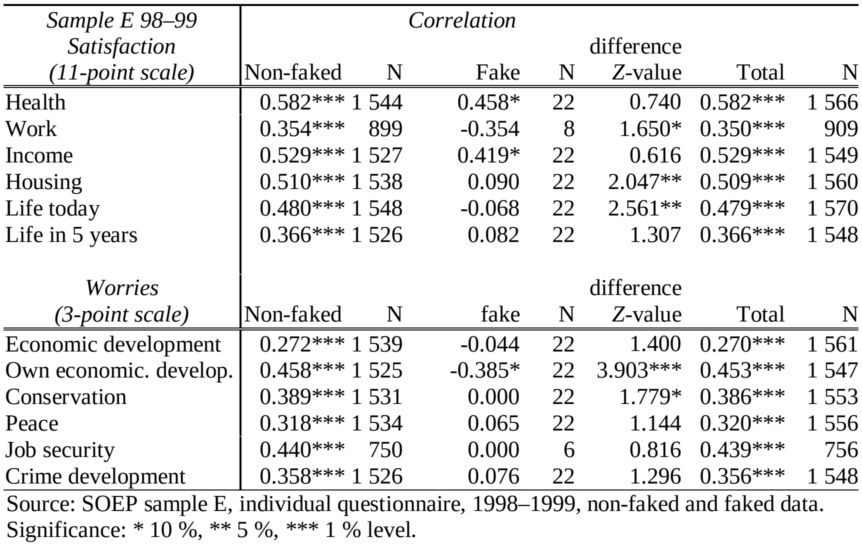
The possible bias due to falsifications is formally similar to the possible bias due to imputation of values in the case of missing data. We can interpret falsifications as a special kind of imputation that depends on an interviewer's assumptions about an unknown respondent's characteristics and opinions (cf. Schnell, 1991). In this section we show some simple equations for calculating the possible bias due to interviewer cheating[9] (following Kalton, 1983, p. 6–10, and extended for our problem). For simplicity, we consider a simple random sample of size n drawn from a population of size N, and we first concentrate on a single variable Y. Let N_{nf} be the number of non-faked interviews and N_f be the number of faked interviews in the population, with $N = N_{nf} + N_f$. The corresponding sample quantities are n_{nf}

[7] We used a Z-test for unpaired equal correlations. We transform the correlations into variables that are approximately normally distributed by using Fisher's Z transformation.

$$Z = \frac{Z_1' - Z_2'}{s_p} = \frac{\frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) - \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)}{\sqrt{1/(n_1-3) + 1/(n_2-3)}} \overset{appr.}{\sim} N(0,1)$$

[8] We find that the cheating interviewer provides consistent values for demographic variables such as gender and year of birth in both waves of sample E (1998 and 1999).

[9] Schnell (1991) also gives equations for the possible bias in samples with faked data. Nevertheless, he makes the implicit assumption that the data for the faked cases is the same as the data for the non-faked part of the sample. He does not explicitly distinguish between the unknown true data and the non-faked data.



Sample E 98–99 *Satisfaction* *(11-point scale)*	*Correlation* Non-faked	 N	 Fake	 N	 difference *Z*-value	 Total	 N
Health	0.582***	1 544	0.458*	22	0.740	0.582***	1 566
Work	0.354***	899	-0.354	8	1.650*	0.350***	909
Income	0.529***	1 527	0.419*	22	0.616	0.529***	1 549
Housing	0.510***	1 538	0.090	22	2.047**	0.509***	1 560
Life today	0.480***	1 548	-0.068	22	2.561**	0.479***	1 570
Life in 5 years	0.366***	1 526	0.082	22	1.307	0.366***	1 548
Worries *(3-point scale)*	Non-faked	N	fake	N	difference *Z*-value	Total	N
Economic development	0.272***	1 539	-0.044	22	1.400	0.270***	1 561
Own economic. develop.	0.458***	1 525	-0.385*	22	3.903***	0.453***	1 547
Conservation	0.389***	1 531	0.000	22	1.779*	0.386***	1 553
Peace	0.318***	1 534	0.065	22	1.144	0.320***	1 556
Job security	0.440***	750	0.000	6	0.816	0.439***	756
Crime development	0.358***	1 526	0.076	22	1.296	0.356***	1 548

Source: SOEP sample E, individual questionnaire, 1998–1999, non-faked and faked data.
Significance: * 10 %, ** 5 %, *** 1 % level.

TABLE 3. Stability coefficients for items of satisfaction and worries in sample E (1998–1999).

and n_f, with $n = n_{nf} + n_f$. The population total is given by $Y = Y_{nf} + Y_t$, and the population mean is given by $M = N_{nf}/N \cdot M_{nf} + N_f/N \cdot M_t$, where Y_{nf} and M_{nf} are the total and mean for non-faked data and Y_t and M_t are the same quantities for the nonrespondents.[10] The corresponding sample quantities are $y = y_{nf} + y_t$ and $m = n_{nf}/n \cdot m_{nf} + n_f/n \cdot m_t$. Cheating interviewers try to impute the missing values of the nonrespondents with faked data in the sample. The sample quantities y_f and m_f are the total and the mean for the faked data. If the faked data are not detected, the sample mean $m_w = n_{nf}/n \cdot m_{nf} + n_f/n \cdot m_f$ contains not the true but the faked quantities for the nonrespondents. This sample mean m_w is used to estimate the population mean M. Its bias is given by $B(m_w) = E(m_w) - M$. The expectation of m_w is

$$E(m_w) = E\left(\frac{N_{nf}}{N}E(M_{nf}) + \frac{N_f}{N}E(M_f)\right) = \frac{N_{nf}}{N}M_{nf} + \frac{N_f}{N}M_f\,.$$

Hence, the bias of the mean m_w is given by

$$\begin{aligned} B(m_w) = M_w - M &= \frac{N_{nf}}{N}M_{nf} + \frac{N_f}{N}M_f - \frac{N_{nf}}{N}M_{nf} - \frac{N_f}{N}M_t \\ &= \frac{N_f}{N}(M_f - M_t)\,. \end{aligned} \tag{1}$$

Equation (1) shows that m_w is approximately unbiased for M if either the proportion of the fakes N_f/N is small or the mean in the faked data is

[10] These ‘nonrespondents’ are the member of the population who were not actually contacted and will be later be assigned faked responses by cheating interviewers.

close to that for the unknown true data M_t. Unfortunately, we have no direct empirical evidence on the magnitude of $(M_f - M_t)$. If we assume that the unknown true values y_t have the same distribution as the known non-faked values y_{nf}, we get a rough estimate for the empirical bias $B(m_w)|(M_t = M_{nf}) = N_f/N(M_f - M_{nf})$. In the case of proportions, the bias is given by

$$B(p_w) = P_w - P = \frac{N_f}{N}(P_f - P_t)\,. \qquad (2)$$

Equation (2) shows that the bias for the proportion cannot be greater than the proportion of the falsified values in the sample. Hence, if there are 3 % fakes in the sample, the maximum bias can be no more than 3 %. Again, if we assume equal distribution for non-faked and true data, we will get an estimate of the empirical bias with $B(p_w)|(P_t = P_{nf}) = N_f/N(P_f - P_{nf})$.

Finally, we consider the effect of cheating on the estimation of variances and covariances. The expectation of the respondent sample variance s_w^2 is $E(s_w^2) = E(S_w^2) = S_w^2$, where

$$S_w^2 = \frac{N_{nf}}{N} S_{nf}^2 + \frac{N_f}{N} S_f^2 + \frac{N_{nf}}{N}\frac{N_f}{N}(M_{nf} - M_f)^2\,.$$

The bias of s_w^2 as an estimator for S^2 is thus $B(s_w^2) = S_w^2 - S^2$, where

$$S^2 = \frac{N_{nf}}{N} S_{nf}^2 + \frac{N_f}{N} S_t^2 + \frac{N_{nf}}{N}\frac{N_f}{N}(M_{nf} - M_t)^2\,.$$

Hence, the bias is

$$B(s_w^2) = \frac{N_f}{N}(S_f^2 - S_t^2) + \frac{N_{nf}}{N}\frac{N_f}{N}\left[(M_{nf} - M_f)^2 - (M_{nf} - M_t)^2\right]\,. \qquad (3)$$

The first term of this bias is comparable to the bias for a mean and the proportion in (1) and (2). The second term reflects the effect of differences in the non-faked and faked mean as well as the true mean on the estimator. Under the assumptions of equal distribution for non-faked and true data so that $S_{nf}^2 = S_t^2$ and $M_{nf} = M_t$, we will get an estimate for the empirical bias with $B(s_w^2)|(S_{nf}^2 = S_t^2, M_{nf} = M_t) = N_f/N(S_f^2 - S_{nf}^2) + N_{nf}/NN_f/N(M_{nf} - M_f)^2$.

For the covariance another variable, x, needs to be introduced. We assume that respondents provide both x and y values. The expectation of the sample covariance s_{xyw} is $E(S_{xyw}) = S_{xyw}$, where

$$S_{xyw} = \frac{N_{nf}}{N} S_{xynf} + \frac{N_f}{N} S_{xyf} + \frac{N_{nf}}{N}\frac{N_f}{N}(M_{xnf} - M_{xf})(M_{ynf} - M_{yf})\,.$$

The bias of s_{xyw} as an estimator of S_{xy} is $B(s_{xyw}) = S_{xyw} - S_{xy}$, where

$$S_{xy} = \frac{N_{nf}}{N} S_{xynf} + \frac{N_f}{N} S_{xyt} + \frac{N_{nf}}{N}\frac{N_f}{N}(M_{xnf} - M_{xt})(M_{ynf} - M_{yt})$$

and the bias is

$$B(s_{xyw}) = \frac{N_{nf}}{N}(S_{xyf} - S_{xyt}) + \frac{N_{nf}}{N}\frac{N_f}{N} \\ [(M_{xnf} - M_{xf})(M_{ynf} - M_{yf}) \\ -(M_{xnf} - M_{xt})(M_{ynf} - M_{yt})] \,. \tag{4}$$

4.2. Empirical bias due to interviewer cheating in SOEP. In this section we present some empirical findings from our analysis of faked and non-faked data. First, we look at some descriptive statistics such as proportions, means, and variances. We analyze only samples A, B, and E because the number of detected fakes in subsample F is too small ($N = 8$). While we are able to give valid values for the maximal possible bias in the case of means and proportions, the given values for the empirical bias are only estimates under the assumption that the distribution of the unknown true data follows the known non-faked data. The only information available for the unknown true data is the area characteristic (Table 2) and the size of the interviewer's assignment. While the assumption of equal distribution is supported for sample E on the basis of these two variables, we found significant differences for sample A. Hence, for sample A the estimation for the empirical bias should be regarded with caution.

Proportions. In the previous section, we demonstrated that the possible bias in the case of proportions cannot be greater than the proportion of falsified values in the sample. The next three tables show proportions and frequencies of some selected variables.

Table 4 shows the breakdown of gender responses in faked and non-faked samples. The last two columns contain information on the estimation of the empirical bias and the possible maximal bias.[11] The results show that we can detect only marginal empirical biases for samples A and E. It can be assumed that it is rather easy for cheating interviewers to reproduce responses like respondents' gender because the distribution is known. Hence, we will take a look at other variables with more categories.

Table 5 shows the distribution of respondents' employment status in samples A and B for faked and non-faked data. This variable has seven categories. Surprisingly, the distribution for the faked sample is quite similar to the non-faked data.[12] The ranking order of the categories corresponds in both data sets and there are only small deviations in the frequency values. Therefore we can expect that the cheating interviewers have an idea of the distribution of the employment status in the entire population and are able to reproduce the frequencies of this variable.

[11] The maximal possible bias is if for example all true respondents are female and all faked respondents are male.

[12] A chi-square test shows that the difference is not significant at a 10 % level (chi-square value is 8.08).

Sample A *1984*					Total		Estim.	
Respondent's gender					observed		emp.	Max.
	Non-Fake	%	Fake	%	data	%	bias	± bias
Male	4 328	47.686	27	45.763	4 355	47.674	- 0.012	0.646
Female	4 748	52.314	32	54.237	4 780	52.326	0.012	0.646
Total	9 076	100.00	59	100.00	9 135	100.00		
Sample E *1998*								
Male	932	48.796	26	55.319	958	48.952	0.156	2.402
Female	978	51.204	21	44.681	999	51.048	-0.156	2.402
Total	1 910	100.00	47	100.00	1 957	100.00		

Source: SOEP, individual questionnaire, sample A, B, and E, non-faked and faked data.

TABLE 4. Proportion of respondent's gender in faked and real samples (individual questionnaire).

Employment status	Non-Fake		Fake		Total	
	N	%	N	%	N	%
Full-time employment	5 724	46.7	57	54.8	5 781	46.8
Reg. part-time employment	661	5.4	9	8.7	670	5.4
Vocational training	439	3.6	4	3.8	443	3.6
Marginal part-time employment	341	2.8	1	1.0	342	2.8
Unemployed, military, community service	474	3.9	5	4.8	479	3.9
Not employed	4 606	37.6	28	26.9	4 634	37.5
Total	12 245	100	104	100	12 349	100

Source: SOEP sample A and B, real and faked data, individual questionnaire.

TABLE 5. Distribution of employment status in sample A and B, 1984 (faked and non-faked data).

MEANS AND VARIANCES. Table 6 shows some means and variances in faked and non-faked data. We have calculated the means for satisfaction items (11-point scale). Under the assumption that the distributions of non-faked and the unknown true data are similar we find that the estimated empirical bias in all cases is rather low and negligible; only half of the differences in means between faked and non-faked data are significant.

CORRELATIONS. In this section we examine the influence of faked data on bivariate statistics such as covariances and correlations. Table 7 shows the correlation between net and gross income as well as between gross income and duration of training (years) in samples A and E. The relationship between gross and net income is trivial, obviously in both non- faked and faked data. However, the connection with 'duration of training' (generated from the variable for schooling and training in years) is more complicated and more adjustments are required. On the basis of human capital theory, we expect a positive correlation and find a significant positive value of 0.367 (sample A) and 0.342 (sample E) in the non-faked data. In the faked data,

Sample			*E*		*1998*			Estim.	
satisfaction	Non-fake		Fake		Total		*F*-Test	emp.	Max.
(11-point scale)	mean	σ^2	mean	σ^2	mean	σ^2	prob.	bias	± bias
Health	7.00	5.681	8.06	1.974	7.02	5.617	0.002***	0.02	0.246
Work	7.17	5.772	7.56	1.949	7.18	5.681	0.411	0.01	0.246
Household	6.75	5.036	6.72	2.606	6.75	4.972	0.944	0.00	0.246
Income	6.45	5.706	7.34	2.229	6.47	5.639	0.011**	0.02	0.246
Housing	7.97	3.991	8.38	0.633	7.98	3.913	0.157	0.01	0.246
Leisure	7.40	5.011	7.87	1.809	7.41	4.938	0.148	0.01	0.246
Products on offer	6.44	7.963	6.55	6.122	6.44	7.915	0.781	0.00	0.246
Public transport	6.40	8.239	7.28	4.161	6.42	8.156	0.038**	0.02	0.246
Environmental sit.	6.77	4.249	7.79	0.432	6.80	4.181	0.001***	0.03	0.246
Living standard	7.36	3.441	8.02	0.934	7.37	3.391	0.015**	0.01	0.246
Life today	7.44	3.001	8.09	0.427	7.46	2.948	0.011**	0.02	0.246
Life in 5 years	7.23	3.836	8.06	0.539	7.25	3.772	0.004***	0.02	0.246

Source: SOEP 1998, sample E, individual questionnaire, non-faked and faked data.
Significance: * 10 %, ** 5 %, *** 1 % level.

Table 6. Means and variances in faked and non-faked data (sample E).

we get partly inconsistent results: In sample A the correlation is 0.470 and not significantly higher than in the non-faked data, and in sample E only a small negative insignificant correlation occurs. Although the amount of fakes in sample E is under 5 % and very small, the impact of the fakes in the overall sample on the correlation is serious, biasing the total positive correlation downward to a value of 0.271.

Linear regressions. Next, we examine the impact of fakes on multivariate statistics such as linear regressions. One of the most important regressions in a socio-economical context is the regression of log gross income. In our equation we use 'age' (in years), 'age squared', 'gender', 'duration of training', and 'working hours per week' as right-hand variables. Table 8 shows the estimated parameters for samples A and E (1998). In the non-faked samples, all coefficients have the expected signs and are significant, the log gross income increases with duration of training, working hours and the age of respondents (proxy for vocational experience), and male respondents have higher incomes than females. The coefficients are reasonable and the overall fit of this model is measured in both samples with adjusted $R^2 = 0.667$ and $R^2 = 0.562$. In the faked data sets (0.6 % of the subsam-

Correlation between gross income	Sample A				Sample E			
and . . .	net income	N	training	N	net income	N	training	N
Non-fake	0.942***	4 216	0.372***	4 345	0.948***	699	0.342***	620
Fake	0.988***	30	0.485***	30	0.924***	27	-0.004	27
Total	0.943***	4 246	0.372***	4 375	0.948***	726	0.271***	747

Source: SOEP sample A (1984) and sample E (1998), individual questionnaire, non-faked and faked data. Significance: * 10 %, ** 5 %, *** 1 % level.

Table 7. Correlations in faked and non-faked data of sample A and E.

ple A and 4.7 % of the subsample E), we find inconsistent results. While in sample A, the estimated parameters are rather close to those in the non-faked data set (with exception of the overestimation of training), we find some differences in sample E. Here, the coefficient for duration of training is only one-third, and the coefficient for age is more than 60 % higher than the coefficient in the non-faked data set. If we leave the faked data in sample E in, we get biased estimates. In the overall sample E, with regard to the non-faked data, the sign of the estimated parameters does not change, but the covariate 'age' is overestimated and 'duration of training' is underestimated. The overall fit is lower than in the non-faked data set, the value for R^2 declines to 0.549.

	Sample A						*Sample E (1998)*					
Regression on log gross income												
	Non-fake		Fake		Total		Non-fake		Fake		Total	
	coeff.	*t*-value	coeff.	*t*-value	coeff.	*t*-value	coeff.	*t*-value	coeff.	*t*-value	coeff.	*t*-value
const.	4.314***	53.87	4.361***	8.04	4.316***	54.07	5.426***	28.88	5.342***	5.16	5.579***	29.46
Age	0.079***	17.27	0.038*	1.64	0.078***	17.27	0.062***	6.21	0,104**	2.35	0.068***	6.67
Age squared	-0.001***	-14.60	-0.001	-0.66	-0.001***	-14.56	-0.001***	-4.70	-0.001**	-2.07	-0.001***	-5.24
Gender (1 = men)	0.271***	17.27	0.189	1.03	0.271***	17.34	0.144***	3.61	0.135**	1.86	0.142***	3.61
Duration of training (years)	0.071***	22.80	0.134**	2.72	0.071***	22.88	0.055***	7.19	0.018	1.88	0.034***	6.27
Empl. (1 = full time)	1.003***	42.47	0.886***	5.54	1.001***	42.58	0.952***	14.72	0.782	4.82	0.953***	14.98
R²	0.667		0.805		0.667		0.562		0.754		0.549	
interviewer (STRATA)	506		3		509		99		2		101	
N (PSU)	4 218		30		4 248		577		25		602	

Source: SOEP sample A (1984) and sample E (1998), individual questionnaire, non-faked and faked data.

TABLE 8. Parameters of the linear regression on log gross income (non-faked and faked data in sample A and sample E).[13]

5. SUMMARY AND CONCLUSIONS

This paper deals with faked interviews in the German Socio-Economic Panel (SOEP), the detection of these fakes, and their impact on survey results. A total of 90 faked household interviews and 184 faked individual interviews were detected mainly by the verification method, almost all of them in the first wave of each subsample. The share of faked data is low in all samples (far less than 1 %) and the maximum is 2.4 % in sample E. Except for the fakes in sample E, faked data was never disseminated within the widely-used SOEP: The fakes were detected before the data were released. However, these fakes are contained in the original data files – which are archived at DIW Berlin – and offer a rich data source for methodological research.

[13] To take cluster effects into account we use the linear regression procedure for complex survey data, implemented in STATA.

We analyze the impact of faked interviews on survey results. We show that the impact of interviewer cheating on proportions cannot be greater than the proportion of the fakes in the sample. Under the assumption that the distribution of the unknown true data follows the known non-faked data, we give estimates of the empirical bias. Overall we observed that the estimated bias for proportions is very small and negligible in SOEP, not only because the share of fakes is low, but because the 'quality' of fakes is high: Interviewers who cheat have successfully reproduced the frequency of particular variables such as 'respondent's gender' and 'employment status'.

Whereas the estimated bias of proportion and means is not noteworthy, we find effects on correlations and regressions in sample E where the share of fakes is higher than in the other samples. We find that some cheating interviewers fail to reproduce the covariance between schooling and gross income as well as the linear regression on the log income. Our empirical results show that the resulting parameters may be seriously biased. Therefore, we find empirical evidence for Schnell's (1991) finding based on simulation results that even small proportions of fake interviews are an important problem in multivariate survey statistics.

Nevertheless, because the number of cases of faked interviews used in our analysis is so small, we have to draw our conclusions with some reservation. Our findings might not apply to other surveys with different quality control methods or different survey management practices.

References

Biemer, P. P., Stokes, S. L. (1989). The optimal design of quality control samples to detect interviewer cheating. *Journal of Official Statistics* **5** 23–39.

Bushery, J. M., Reichert, J. W., Albright, K. A., Rossiter, J. C. (1999). Using date and time stamps to detect interviewer falsification. Proceedings of the American Statistical Association, Survey Methods Section, 316–320. http://www.amstat.org/sections/srms/Proceedings/.

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

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 Methods Section, 820–824. http://www.amstat.org/sections/srms/Proceedings/.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Institute for Social Research, Ann Arbor.

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 für Soziologie* **20** 25–35.

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** 163–175.

Schreiner, I., Pennie, K., Newbrough, J. (1988). Interviewer falsification in Census Bureau surveys. Proceedings of the American Statistical Association, Survey Methods Section, 491–496. http://www.amstat.org/sections/srms/Proceedings/.

Stokes, S. L., Jones, P. (1989). Evaluation of the interviewer quality control procedure for the post-enumeration survey. Proceedings of the American Statistical Association, Survey Methods Section, 696–198. http://www.amstat.org/sections/srms/Proceedings/.

Wagner, G. G., Burkhauser, R. V., Behringer, F. (1993). The English language public use file of the German Socio-Economic Panel. *The Journal of Human Resources* **28** 429–433.

Joerg-Peter Schraepler
DIW Berlin
(and Ruhr Universität Bochum)
Königin-Luise-Str. 5
14191 Berlin
jschraepler@diw.de

Gert G. Wagner
DIW Berlin
(and Technische Universität Berlin)
Königin-Luise-Str. 5
14191 Berlin
gwagner@diw.de