Wagner, Gert G.; Schraepler, Joerg-Peter

**Working Paper**

# Identification, Characteristics and Impact of Faked Interviews in Surveys : An analysis by means of genuine fakes in the raw data of SOEP

DIW Discussion Papers, No. 392

**Provided in Cooperation with:**
German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Wagner, Gert G.; Schraepler, Joerg-Peter (2003) : Identification, Characteristics and Impact of Faked Interviews in Surveys : An analysis by means of genuine fakes in the raw data of SOEP, DIW Discussion Papers, No. 392, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/18154

DIW Berlin

German Institute
for Economic Research

Discussion Papers

392

Joerg-Peter Schraepler
Gert G. Wagner

# Identification, Characteristics and Impact of Faked Interviews in Surveys – An analysis by means of genuine fakes in the raw data of SOEP

Berlin, Dezember 2003

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

DIW Berlin

German Institute
for Economic Research

Königin-Luise-Str. 5
14195 Berlin,
Germany

Phone +49-30-897 89-0
Fax +49-30-897 89-200

www.diw.de

ISSN 1619-4535

# Identification, Characteristics and Impact of Faked Interviews in Surveys

## - An analysis by means of genuine fakes in the raw data of SOEP

by Joerg-Peter Schraepler* and Gert G. Wagner**

***Abstract***

To the best of our knowledge, most of the few methodological studies which analyze the impact of faked interviews on survey results are based on "artificial fakes" generated by project students in a "laboratory environment". In contrast, panel data provide a unique opportunity to identify data which are actually faked by interviewers. By comparing data of two waves, unequivocal fakes are easily identifiable. However, in most surveys there is no second wave because they have a pure cross-sectional nature. In search of a method which does not need two waves of data we test an unconventional benchmark called Benford's Law, which is used by several accountants to discover frauds. Our preliminary results let us conclude that Benford´s Law might be not an efficient method for detecting faked data, but it might be a new instrument for quality control of the interviewing process

The raw data of the German Socio-Economic Panel Study (SOEP) provide a rich source of faked interviews because it is built on several sub-samples. However, because interviewers know that panel respondents will be interviewed again over the course of time, clever interviewers will not fake panel interviews. In fact, in raw data of SOEP the share is about only 0,5 percent of all records. The fakes are used for an analysis of the potential impact of non detected fakes on survey results. The major result is that the faked records has no impact on the mean and the proportions. But in very rare, exceptional cases there may be a bias in estimates of correlations and regression coefficients if fakes would not be detected. One should note that – except for some fakes in the first two waves of sample E – faked data were never disseminated within the widely-used SOEP. The fakes were detected before the data were released.

Key words: Benford's law, cheating; curbstoning; faked interviews, quality control, SOEP.

*JEL Classification: C 8, C 4*

---

* Ruhr University Bochum and DIW Berlin, 14191 Berlin, jschraepler@diw.de

** DIW Berlin and Berlin University of Technology (TUB), 14191 Berlin, gwagner@diw.de

# 1 Introduction*

In any survey in which the data are collected by personal interviews there is a danger of cheating by interviewers, or that some interviewers may fabricate data.[1] 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. Bureau of the Census refers to this practice as "curbstoning", thus named because a census taker "stands at the curb" and guesses the number of residents in a building or house without ever entering. Interviewers who do this are called curbstoners[2] (cf. Moore and Marquis 1996).

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, such as conducting an interview with someone who is easily reachable and willing to participate in 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 the telephone when a personal visit is required.

In our paper we deal only with the first form of cheating, the fabrication of an entire interview. This is sometimes called "curbstoning" in the literature.

We focus on fabricated data in the German Socio-Economic Panel (SOEP) which contains unique "genuine" fakes because faked data of wave 1 which were detected after the fieldwork of wave 2 were already delivered from the fieldwork organization to DIW Berlin which hosts the SOEP study. These raw data were kept by DIW Berlin and can be analyzed. Whereas in usual surveys the fieldwork organization never detect fakes or it deletes fakes from the raw data before the deliver them to a client.

We will start our analyses by providing some hints about the quality control procedures which were used by the fieldwork organization to detect the fakes in SOEP. These verification methods are a subject of the few other studies which deal with fakes in surveys (c.f. Biemer/Stokes 1989; Stokes/Jones 1989; Bushery et al. 1999). Therefore in our paper we will try to go two steps further and we discuss an additional procedures which may be useful for detecting cheating behavior in panel surveys if the usual quality control is passed. Because in cross-sectional surveys there is no second wave we were in search of a method which does not need two waves of data. We found and tested an unconventional benchmark called Benford's Law, which is used by several accountants to discover frauds. In our last section we examine the possible and empirical impact of cheating on survey results.

From a SOEP user point of view it is worthwhile to mention that due to the quality control procedure of the fieldwork organization never faked data were available for a significant

* We are grateful to participants of the Workshop on "Item Non-response and Data Quality on Large Social Surveys", at University of Basel for good critisizm and comments, especially to Rainer Schnell and our brilliant discussant Regina Riphahn. The usual disclaimer applies.

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

[2] Curbstoning is a term that originated with 18th-century census-taking. This term was coined when it was discovered that some interviewers simply filled out interview schedules without even contacting a respondent.

analyses. In the data set which is available for analyses faked data are deleted. There was only one exception in subsample E. Faked data on 11 households were in the data set for two waves and were taken out with the release of wave 3 of subsample E.

## 2 Previous results on cheating behavior

Compared to other methodological topics of statistics, there are only few studies dealing with cheating by interviewers known 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 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, Pennie, and Newbrough 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 re-interviews[3] and the majority (79%) was determined to have fabricated interviews.[4] Their results provide evidence that the shorter the length of service, the more likely 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 deal with the "quality" of faked interviews and the impact of fabricated data on substantive analysis. Reuband (1990) shows that students are able to reproduce data in fictive interviews using given demographic variables of real respondents (study on Germany).

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 fabricated by sociology students and university colleagues. He analyses the quality of the fabricated data and the robustness of substantive empirical results by comparing the German General Social Survey with the substituted faked data. His main result is that univariate statistics such as proportions, means and variances are relatively robust against typical amounts of fabricated data in surveys (less than 5%). Nevertheless he also found some minor effects on multivariate statistics such as multiple regressions. Moreover, using simulations he shows that higher proportions of fabricated data in surveys will have a serious impact on multivariate statistics

[3] Reinterviews are an effective component of the interviewer quality control program of the U.S. Census Bureau.
[4] The second highest type of falsification (18.5 percent) was deliberately misclassifying units as vacant when they were occupied (Schreiner/Pennie/Newbrough 1988). In an update of the database including information on all confirmed cases of falsification from 1982 to 1992, 305 cases of falsification were found (Cantwell, Bushery and Biemer 1992).

and data quality. In our paper we intend to prove his predictions based on simulation results with real faked data.

In the German ALLBUS in 1994 the ADM design was replaced with a new sampling design, which offers the opportunity to systematically check that the interviews (N = 3505) are performed correctly. The interviewers are given the address and the names of the respondent directly. In six percent of the cases irregularities were detected; half of them turned out to be faked by the interviewers (Koch 1995). These fabricated data (n = 45) are found *after* the routine monitoring by the data collection institute via the postcard method, which detected fifteen faked interviews in this survey. Another finding was that interviewers who cheat are mainly younger persons with higher education (*Abitur*) and with a relatively high workload (number of interviews). The interviewer characteristics of cheating interviewers are known in the SOEP (Schräpler/Wagner 2001). Therefore we are able to compare them with the characteristics found in the ALLBUS.

A rare debacle caused by faked interviews is mentioned by Diekmann (2002). In the German city Rostock a traffic study about drivers was carried out by means of 600 face-to-face interviews. Eighty cases were later re-contacted for another study, which showed that sixteen of the former interviews were completely or partly fabricated by the interviewer. If we extrapolate this to the whole sample, that amounts to a share of 20% fakes.

## 3 Quality control and detecting cheating interviewers in the SOEP

In contrast to cross-sectional surveys, curbstoning is extremely difficult in complex long-term panel studies such as the SOEP (German Socio-Economic Panel Study) because the respondent is interviewed face to face every year, and because a consistency check between waves shows irregularities immediately.[5] Hence we can assume that fabricated data will be a problem mainly in the first wave and will be detected quickly after conducting a second wave. The results of the quality control show that this was clearly the case in the SOEP. However, because the fieldwork organization does not wait with the release of the raw data to DIW Berlin, which makes the data "user friendly" and disseminate them to a worldwide "user community", the DIW Berlin get faked data which can be analysed (but it was never done up to now). Due to the fact that the process of making the raw data user-friendly takes a while the detection of fakes after wave 2 (by the fieldwork organization) is early enough to prevent the dissemination of those fakes. There was one exception only: in subsample E faked data for 11 households were detected not until wave 3, so the data for waves 1 and 2 were already released. They were taken out of the data set (for all waves) with the release of wave 3 of sample E.

[5] This argument is relative and depends on the time lapsed between interviews for the same housing unit and the data collection method. Schreiner et al. (1988) investigates falsifications in the Current Population Survey (CPS) and the National Crime Survey (NCS). The CPS is a panel-type survey in which the same housing unit is interviewed monthly, whereas in the NCS interviews for the same housing unit are conducted at six-month intervals and the data collected concerns incidents of crime occurring over the previous six months. In the CPS, continuing households can for the most part be interviewed by telephone and without any personal visits. This reduces the risk of an interviewer being caught faking data. Schreiner at al. found that especially experienced interviewers have tried to fabricate data in the CPS. They argued that the data collected often can be "imputed" correctly from a past month to the current month. In the case of the NCS they found the opposite: mainly interviewers with less than one year of service had tried to fabricate data. In the SOEP the same housing unit is interviewed yearly 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.

### *3.1 Verification methods*

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, Pennie, and Newbrough 1989). In this sense the German Socio-Economic Panel (SOEP) provides a unique opportunity to identify fabricated data. Falsifications are detected in several ways:

1. Most fakes are identified easily by comparing data of two waves. If data deviate considerably from the data of the previous year(s), the interview control department contacts the respective households by phone and the household members are asked to verify the data.
2. If there is a change in interviewer in the following wave, in the case of falsified data the new interviewer cannot confirm the composition of the household as recorded in the address protocol.
3. 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, the intended respondent is likely to contact the fieldwork organization, which then becomes aware of the falsified interviews. If the recorded address does not exist, the interviewer control department is informed.
4. Due to problems with curbstoners in sample E of SOEP (1998), for sample F (2000) all households were re-contacted after interviewing and asked to verify the household composition.

Additional control routines are implemented to further secure the quality of the fieldwork. The fieldwork organization uses mainly experienced interviewers for the SOEP project. The average length of service in the first wave is approximately five years.

The SOEP consists of several samples. Fabricated data were always found in the first wave of each sample (with the exception of the East German sample C and the small sample D). Nevertheless, 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 fabricated data. The first wave of samples A and B contains only 0.6 and 1.5% fabricated data, respectively, and the first wave of sample E contains about 2% faked household interviews. In the following wave approximately 1% of fabricated data was identified in sample E. In the first wave of sample F1 the cheating was lowest: only 0.1% of the interviews were detected as fabricated.

**Table 1: Proportion of detected fabricated raw data in the SOEP**

| | *Household interview* | | | *Personal interview* | | |
|---|---|---|---|---|---|---|
| *Sample* | *Valid Cases* | *Fabricated cases* | *Fakes in percent of total cases* | *Valid cases* | *Fabricated 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

### *3.1.1 Area and interviewer characteristics for detected fabricated data in the SOEP*

Because Biemer and Stokes (1989, p.25) find that in the two large demographic surveys cheating behavior differed between urban and rural areas we examine these kind of differences. In the United States, the CPS and the NCS, 87% of the interviews in urban areas were falsified, compared to only 13% in rural areas. Table 2 shows the frequency of fabricated household interviews in Sample A, B, E and F of the SOEP by number of residents in the area. The proportion of falsification in cities (≥ 100,000 residents) is 41.1%, only slightly higher than in smaller areas with 20,000 - 100,000 residents (25.6%) and areas below 20,000 residents (33.3%). Therefore approximately 60% of the fakes are in smaller areas, which indicate that there is no strong area effect.

**Table 2: Distribution of fakes by area characteristic**

| Number of Residents in Area | Sample A+B | Sample E | Sample F | Total | Percent |
|---|---|---|---|---|---|
| ≥ 100,000 | 25 | 12 | - | 37 | 41.1 |
| 20 – 100,000 | 1 | 22 | - | 23 | 25.6 |
| ≤ 20,000 | 22 | - | 8 | 30 | 33.3 |
| Overall | 48 | 34 | 8 | 90 | 100.0 |

Source: SOEP Sample A, B, E and F, household questionnaire

Only very little is known about the characteristics of interviewers who cheat. Koch (1995, 97) shows that younger interviewers with a higher education level have more inconsistencies in their interviews than others. Table 3 lists some characteristics in the case of the SOEP. All interviewers who fabricated data (N = 9) are middle-aged males. We find no education effects; cheating interviewers may have a university degree or only a primary school education. 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 40 personal interviews (not shown in the table).

On average, in SOEP cheating interviewers falsified more than 75% of their assignments in the SOEP. This finding is in line with the results of Hood and Bushery (1997), who show that cheating interviewers usually falsify a relatively high proportion of their assignments. In the SOEP these interviewers are not necessarily new to service, but each of them was working on this panel study for the first time. We can assume that they were not aware of the effectiveness of the quality control in this panel study and of the fact that fakes in this design are easily identifiable.

**Table 3: Characteristics of interviewers with fabricated data**

| Interviewer characteristics | Sample A + B | Sample E + F | Overall |
|---|---|---|---|
| *Gender* | | | |
| - male | 4 | 4 | 8 |
| - female | - | - | - |
| - not known | - | 1 | 1 |
| | | | |
| *Education* | | | |
| - primary school | - | 2 | 2 |
| - secondary school | 2 | 1 | 3 |
| - university | 2 | 1 | 3 |
| - not known | - | 1 | 1 |
| | | | |
| Age (mean) | 44.8 | 44.5 | 44.6 |
| | | | |
| *Years in service* | | | |
| <= 2 | 1 | 3 | 4 |
| > = 3 | 3 | 1 | 4 |
| - not known | - | 1 | 1 |
| | | | |
| *Number of interviews:* | | | |
| - overall HH interviews (mean) | 18.3 | 13.5 | 15.9 |
| - faked HH interviews (mean) | 12.0 | 10.5 | 11.3 |
| - percentage of faked HH interv. | 65.6% | 77.8% | 71.1% |
| - overall personal interv. (mean) | 37.0 | 23.3 | 30.1 |
| - faked personal interv. (mean) | 26.0 | 20.0 | 23.0 |
| - percentage of faked personal int. | 70.3% | 85.8% | 76.4% |
| | | | |
| *Occupation (main job)* | | | |
| -part-time | 2 | - | 2 |
| -full-time | 2 | 3 | 5 |
| -student | - | 1 | 1 |
| - not known | - | 1 | 1 |

Source: SOEP – Interviewer dataset 1984 – 2000

### *3.2 Stability check: weak correlations between consecutively waves*

In most cases cheating is detected instantly after the first wave of a subsample in the SOEP; however, one interviewer was able to fabricate data for the first two waves in sample E. Due to the fact that 22 records were faked - despite the very low number of cases of cheating interviewers - we are able to investigate whether the true and faked answers are consistent from wave to wave or whether there is a difference between the real and the faked data. Table 4 shows the stability coefficients (correlations) for several satisfaction items and items about worries. In the true 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.

Although the stabilities are not given in the fabricated data, we can also recognize that the correlation will not be seriously biased in the total sample. Furthermore, 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).

**Table 4: Stability coefficients for items of satisfaction and worries in Sample E (1998 - 1999)**

| *Sample E 98 - 99* | *Correlation* | | | | | |
|---|---|---|---|---|---|---|
| *Satisfaction* | | | | | | |
| *(11-point scale)* | *true* | *N* | *fake* | *N* | *total* | *N* |
| Health | 0.582*** | 1544 | 0.458* | 22 | 0.582*** | 1566 |
| Work | 0.354*** | 899 | -0.354 | 8 | 0.350*** | 909 |
| Income | 0.529*** | 1527 | 0.419* | 22 | 0.529*** | 1549 |
| Housing | 0.510*** | 1538 | 0.090 | 22 | 0.509*** | 1560 |
| Leisure | 0.459*** | 1538 | 0.055 | 22 | 0.459*** | 1560 |
| Products on offer | 0.607*** | 1542 | 0.122 | 22 | 0.604*** | 1564 |
| Environmental sit. | 0.380*** | 1546 | 0.186 | 22 | 0.380*** | 1568 |
| Living standard | 0.503*** | 1547 | 0.124 | 22 | 0.503*** | 1569 |
| Life today | 0.480*** | 1548 | -0.068 | 22 | 0.479*** | 1570 |
| Life in 5 years | 0.366*** | 1526 | 0.082 | 22 | 0.366*** | 1548 |
| *Worries* | | | | | | |
| *(3-point scale)* | *true* | *N* | *fake* | *N* | *total* | *N* |
| Economical development | 0.272*** | 1539 | -0.044 | 22 | 0.270*** | 1561 |
| Own economic. develp. | 0.458*** | 1525 | -0.385* | 22 | 0.453*** | 1547 |
| Conservation | 0.389*** | 1531 | 0.000 | 22 | 0.386*** | 1553 |
| Peace | 0.318*** | 1534 | 0.065 | 22 | 0.320*** | 1556 |
| Job security | 0.440*** | 750 | 0.000 | 6 | 0.439*** | 756 |
| Crime development | 0.358*** | 1526 | 0.076 | 22 | 0.356*** | 1548 |

Source: SOEP, Sample E, individual questionnaire, 1998-1999, true and faked data

Apparently, the cheating interviewer pays attention to the stability of demographic variables but not to that of subjective indicators. The satisfaction items and items about worries are always a part of the SOEP and we can use them to calculate the average correlation from wave to wave each "interviewer cluster". Because interviewers ask usually several respondents in a survey, one can build so-called "interviewer clusters" where each interviewer cluster contains all interviews from a particular interviewer. Therefore the number of clusters corresponds with the number of interviewers in a survey. For example sample E (1998) is carried out by 107 interviewers who build, of course, 107 clusters. Our research interest is to identify interviewer clusters which contain fabricated data. We have shown in table 3 that the interviewers fake more than 85% of their assignment in sample E which means that the

percentage of true data in clusters from cheating interviewers is rather low. Therefore we can calculate statistical measurements such as means or correlations for each cluster in order to identify outliers. The idea is that interviewers who fake will "create" other "structures" and "patterns" of data than honest interviewers do.

Figure 1 shows the distribution[6] of the average correlation within each cluster. These values are the calculated means of the correlations of each satisfaction and worry item with the corresponding item in the following wave. The distribution follows a nearly normal density function with a mean value of r = 0.42 (0.33 for worries). This positive correlation indicates that overall the measures of satisfaction (12 items) and worries (6 items) are relatively stable from wave to wave within each interviewer cluster (n = 107). But the average stability coefficient for a cheating interviewer has a rare low value of r = 0.08 (for worries r = 0.009). Only two of the 107 interviewers have a stability value lower than 0.1 in cases of both the satisfaction items and the worry items. Of course, a high deviation from the average is not sufficient to indicate a falsification. But in the future we can use this kind of deviation together with other indicators as a way of detecting fabrications and irregularities in the entire data set after quality control is past.

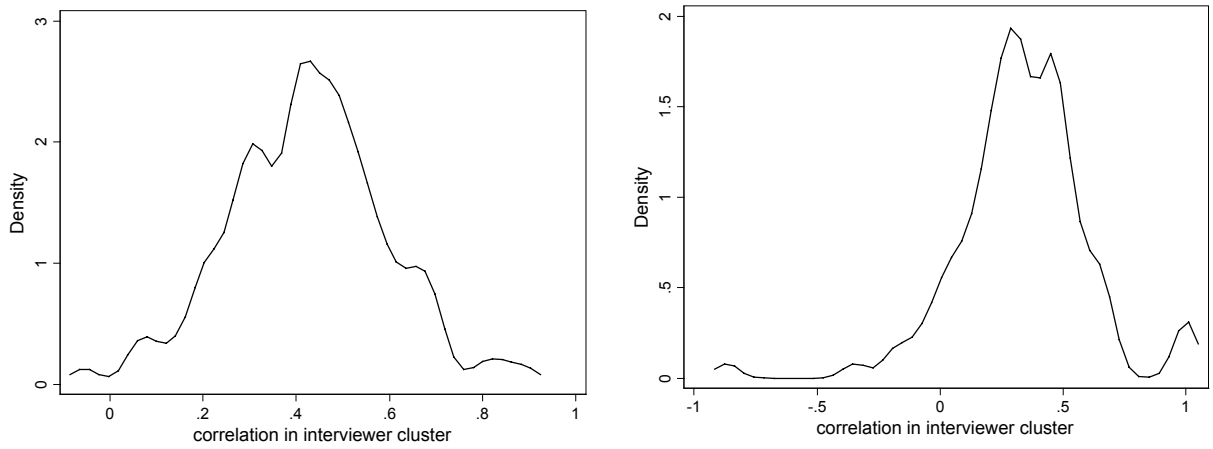


**Figure 1: Distribution of the mean value of the stability coefficient within each interviewer cluster calculated on the basis of 12 satisfaction items (left) and 6 worry items (right), Sample E, individual questionnaire, 1998-1999**

### *3.3 Fraud detection using Benford's Law*

Besides the "conventional" tests of stability and consistence, an unconventional benchmark called Benford's Law has recently been used by several accountants to detect frauds. Some social researchers have proposed using this method for survey data as well (Diekmann, 2002). In this section we test the predictive power of Benford's Law.

[6] We use a kernel density estimation method with a Gaussian kernel.

Benford's Law[7] is an empirical "law" which states that in many tables of numerical data, the significant digits are not uniformly distributed as might be expected, but rather obey a certain logarithmic probability distribution (Hill 1996). The leading significant (non-zero) digit obeys the law

$$\text{Prob(first significant digit = d)} = \log_{10}\left(1+\frac{1}{d}\right), \quad d = 1,2,...,9.$$

Hence, a number chosen at random has leading significant digit $d = 1$ with probability 0.301, a leading digit $d = 2$ with probability 0.176 and so on monotonically down to probability 0.046 for leading digit $d = 9$. The general law for second and higher significant digits and their joint distribution is (Hill 1995, 1999):

$$\text{Prob}(D_1 = d_1,...,D_k = d_k) = \log_{10}\left[1+\left(\sum_{i=1}^{k} d_i \times 10^{k-i}\right)^{-1}\right]$$

where $d_1 \in \{1,2,...,9\}$ and $d_j \in \{0,1,2,...,9\}, j = 2,...,k$ . Therefore the joint probability $\text{Prob}(D_1 = 1, D_2 = 5, D_3 = 2) = \log_{10}(1 + (152)^{-1}) \approx 0.0028$.

The first step towards explaining this relationship was taken in 1961 by the mathematician Roger Pinkham (Pinkham 1961). He argues that if there is a law of digit frequencies, it should be universal and "scale invariant". This means that if we multiply all our numbers by an arbitrary constant, then the distribution of first-digit frequencies should remain unchanged. Pinkham gives the proof that if a law of digit frequencies is invariant under changes of scale (e.g. dollars in euros) then it has to be Benford's Law.

A plausible theoretical explanation for the appearance of this logarithmic distribution is the random-samples-from-random-distribution theorem by the mathematician 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, 360). If Hill's theorem is correct, this means that the digits derived from a random mix of different sources from census data to stock market prices should follow Benford's Law.

There is evidence that many classes of true data sets follow Benford's Law. It has been found the digits of many sets of financial data are following the Benford distribution, including income tax and stock exchange data, corporate disbursements and sales figures, demographics and scientific data (e.g. Nigrini 1999), as well as numbers gleaned from newspaper articles (Benford 1938; Hill 1999). Stock prices may seem to be a single distribution, but their value

[7] According to Hill (1999) in 1881, the astronomer Newcomb (Newcomb 1881) explained that his discovery of the significant-digit law was motivated by an observation that the pages of a book of logarithms were dirtiest in the beginning and progressively cleaner throughout. Nevertheless the law is named for Dr. Frank Benford, a physicist who had made the same observation. In 1938 he embarked on a mathematical analysis of 20,229 sets of numbers, including such wildly disparate categories as the areas of rivers, baseball statistics, numbers in magazine articles and street addresses. He found that all these seemingly unrelated sets of numbers followed the same first-digit probability pattern. In all cases the number 1 turned up as the first digit about 30 percent of the time, more often than any other. Benford derived a formula to predict the frequency of numbers found in many categories of statistics.

actually stems from many measurements (salaries, the cost of raw material and labour) and so it is expected that they will follow Benford's Law in the long run.

In the case of companies in the stock market, which represent all stages of growth, Nigrini gives an additional intuitive explanation. We can consider a growing company with a market value of 100 million €. For the value to reach 200 million, € the company must double its value. For it to increase from 200 million € to 300 million € it must increase only 50%, and for it to increase from 900 million € to 1000 million € it must increase by just 11%. Moreover, for it to increase from 1000 million € to 2000 million € it must again double. Hence a growing company spends longer with a "1" as the first digit of its market capitalisation than it does with any other number. The persistence of a 1 as a first digit will occur with any phenomenon that has a constant or erratic growth rate (Nigrini 1999).

On the other hand, truly random numbers do not confirm to Benford's Law, because the proportion of leading digits in such numbers are, by definition, equal. Those data sets most likely follow to Benford's Law have numbers which do not contain a built-in maximum and describe the sizes of similar phenomena (Nigrini 1999). Assigned numbers, such as Social Security numbers or bank accounts, will not conform to it. Furthermore deviations from the law's prediction can be caused by merely rounding numbers up and down. Moreover, the sample of numbers should be large enough to give the predicted proportions a chance to assert themselves (Pinkham 1961) and the sets of numbers should essentially be subsets of a larger series and not just huge chunks of such series.

Recently Benford's Law has been used to determine the normal level of number duplication in data sets, which in turn makes it possible to identify abnormal digit and number occurrence. Accountants and auditors have begun to apply Benford's law to corporate accounting to discover number-pattern anomalies and frauds. Nigrini found that true tax data have a close fit to Benford and there is substantial evidence that in most fabricated tax data the significant digits are not close to Benford. Usually the faked data have conspicuous patterns and do not follow the expected distribution. Nigrini has used a goodness-of-fit-to-Benford test and successfully identified fraudulent financial data.

### *3.3.1 Using Benford's Law for Survey Data*

An interesting point for survey researchers is whether this logarithmic distribution called Benford´s Law can also be used to identify fabricated data in surveys. Hence the main question is whether survey data follow Benford's Law. Unlike financial data, many variables in these databases are dichotomous or categorical (like gender, marital status and occupation) or are assigned numbers like household numbers.

Figure 2 shows the first-digit distribution of the true and faked data derived from all variables in the individual questionnaire in samples A and B of wave 1.[8] We can recognize that there is a higher proportion of 1's and 2's and a lower proportion of the higher digits in both the true and the faked data than in Benford's distribution. Although the deviation seems to be a little bit higher in the faked data, especially for the numbers 2 and 3, we can't really say that detecting frauds in the overall data is possible with Benford's distribution. Therefore in the next step we will restrict our data to one item which contains metric units as Benford requires.

[8] For this figure we take all metric and non-metric variables into account except the assigned household and personal number as well as the interviewer numbers and the assigned negative values (-1, -2, -3) for missing data. (There are no other negative values in the SOEP.)

Figure 3 shows the first-digit distribution for the item "length of drive to work" measured in km. We can see that in this case the true data set in Sample E follows Benford's distribution rather well and that there is a high deviation in the faked data set for the proportion for number 3. Nevertheless, the reason for this deviation in the faked data may be that the faked data set is too small (n=27).

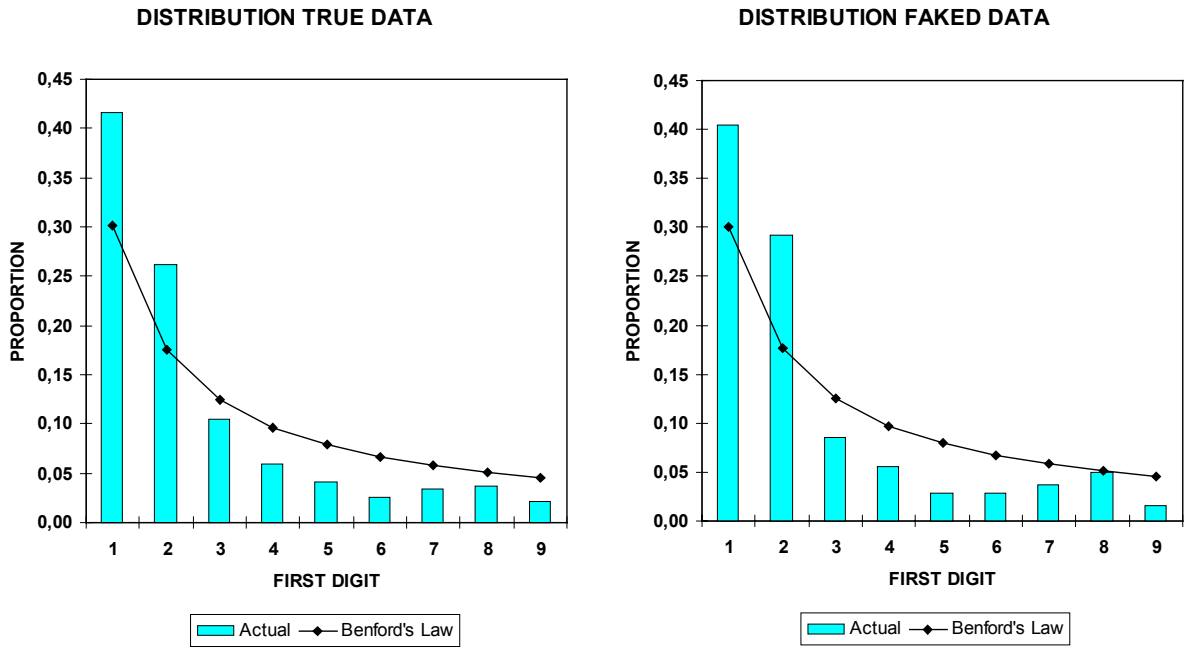


**Figure 2: First-digit distribution of all numbers in individual questionnaire, Sample A + B, 1984, in the cases of true data (n = 990189) and faked data (n = 7603)**

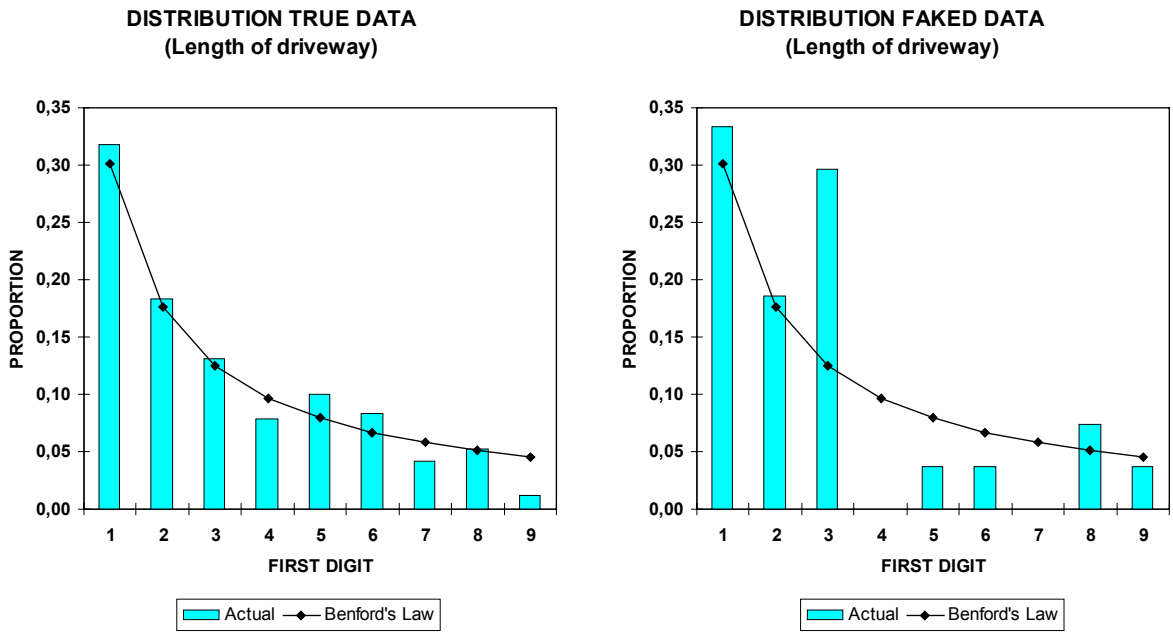


**Figure 3: First-digit distribution of all numbers in individual questionnaire Sample E, 1998, for the item "length of drive" in the cases of true data (n=894) and faked data (n=27)**

In the past Nigrini has successfully applied Benford's Law to tax data and identified fraudulent data. Therefore we will analyse survey data which contains net income and tax data. Figure 4 shows the first- and second-digit distribution of all numbers in samples A and B from waves 1 – 12 for the items "net income" and "taxes".[9] We can observe that in fact the first-digit distribution follows the logarithmic distribution of Benford closely. In the case of the second-digit distribution we recognize a very high proportion of zeros, probably caused by respondents' rounding behaviour (cf. Schräpler 1999).

[9] We find that the gross income in the SOEP does not follow Benford; the proportion of the number 2 is higher than the proportion of number 1.

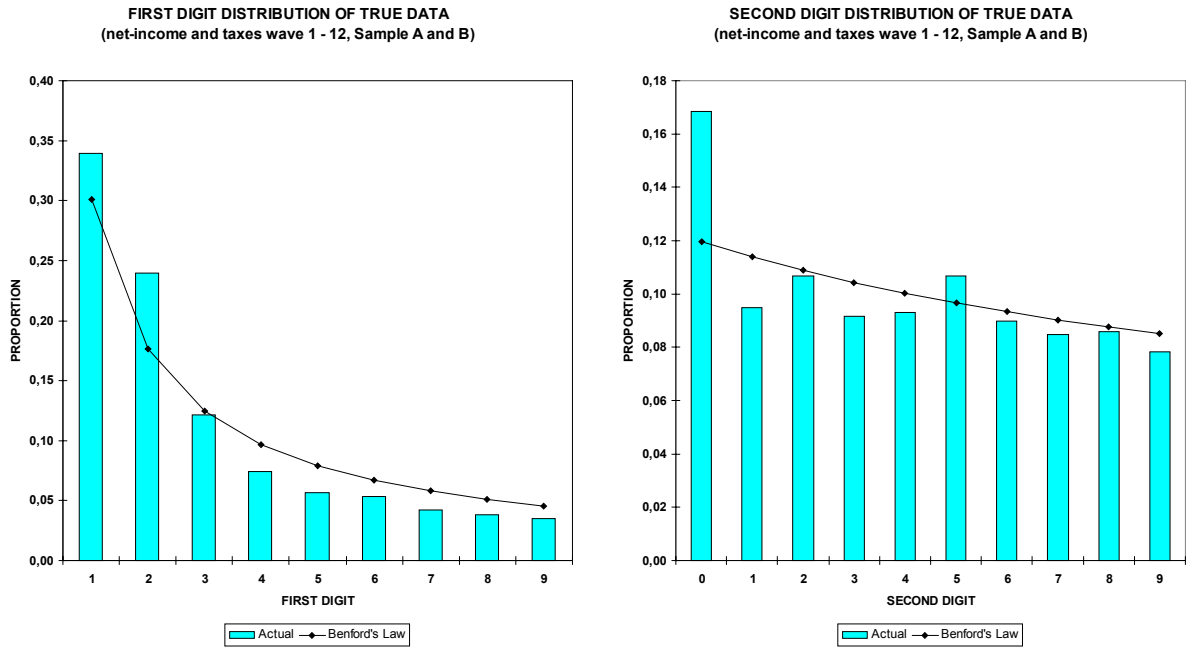


**Figure 4: First- and second-digit distribution of all numbers in individual questionnaire Sample A+B, waves 1 - 12, for the items "net income" and "taxes" in the case of true data (n=91623)**

In Figure 5 we use only wave 1 in samples A and B and compare the distribution in the true and faked data set. Again, we find that the deviation is larger in the fabricated data than in the true dataset, especially the proportions for the numbers 1 and 2. But this finding may be still the result of the small sample size of the faked data set (n = 66).

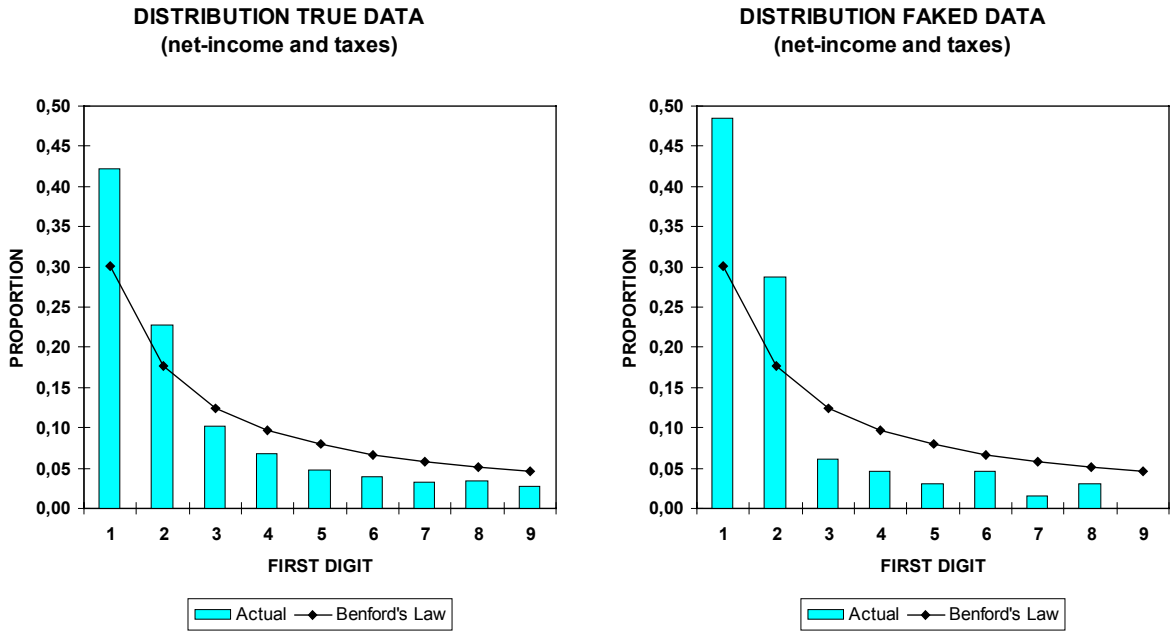


**Figure 5: First-digit distribution of all numbers in individual questionnaire Sample A+B, 1984, for the items "net income" and "taxes" in the cases of true data (n=8198) and faked data (n=66)**

### 3.3.2 Identification of fabricated data by means of Benford's Law

In the last section we could show that for some selected continuous variables, the first-digit distribution in fact follows Benford's logarithmic distribution closely if the sample size is large enough. Now we are going to check whether it is possible to detect cases with fabricated data using Benford's Law. We have shown that the interviewers fabricate a large proportion of their assignment. Therefore it gives more statistical power if we analyse whole clusters of interviews per interviewer ("interviewer cluster") rather than single questionnaires. If real survey data follows the logarithmic distribution and fabricated survey data not, we should be able to identify these clusters of fabricated interviews and to test them for significance.

For our analysis we only use continuously variables like income, taxes or other given monetary values and some items about measurements like distances. All non-metric variables and assigned numbers are excluded. Overall we are able to use 34 sample variables in samples A and B and 73 in sample E. Again, we count the first digits to get the digit distribution in each interviewer cluster. Then we calculate a fit measurement like the chi-square value for each cluster.

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

where $n_i$ is the number of first digits in interviewer cluster $i$, $h_{di}$ is the observed proportion of digit $d = 1,\ldots,9$ in interviewercluster $i$ and $h_{bd}$ is the proportion of digit $d$ under Benford's distribution.

For demonstration Figure 6 shows the distribution for a single selected cluster which contains true data and an interviewer cluster with fabricated data. We can recognize that the fit is much better in the case of the cluster with valid data. But these are only two selected cases. Next we will get an overview of all clusters in the data.

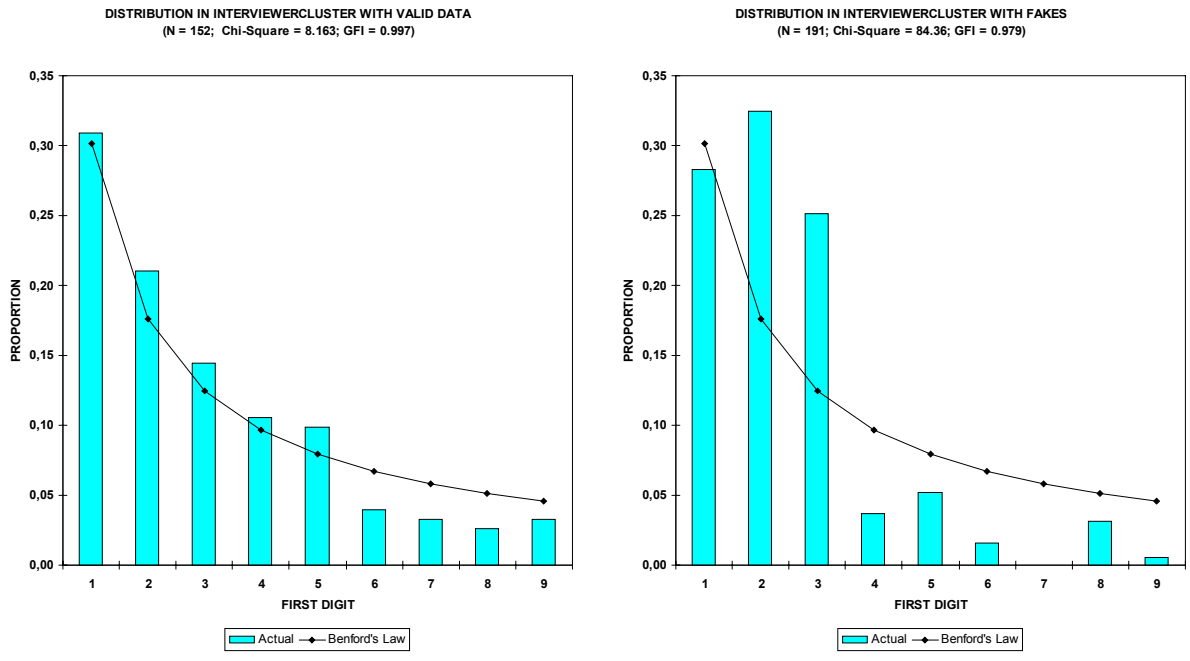


**Figure 6: Two selected interviewer clusters: one cluster with "good" fit which contains only valid data and one with "bad" fit which contains faked data, Sample A and B, individual questionnaire, 1984**

Unfortunately the chi-square value depends on the sample size; however, for reasons of comparison it could be useful to choose an alternative that is independent of the size of the cluster. One possibility is to use a measurement which relates to the worst possible fit. This is the case if all digits in one cluster have the unlikeliest value, the digit 9. We define this goodness of fit (GFI) measurement with

$$GFI = 1 - \frac{\chi_i^2}{\chi_0^2} \quad \text{where} \quad i = 1,..,n$$

the index *i* indicates the interviewer-cluster and $\chi_0^2$ is the chi-square value for the distribution with the worst fit to Benford´s Law. The range is from 0 to 1, where the value 1 indicates an exact Benford distribution and values over 0.99 indicate a very close fit.[10]

The scatter plots in Figure 7 show the chi-square values and the GFI values for all interviewer clusters in samples A, B and E. Lower chi-square values indicate a good fit in each cluster, whereas higher values show that there are larger differences between both distributions. Interviewer clusters with fabricated data are marked with a black circle.

Overall we can recognize in the upper sections that the chi-square values increase with the sample size because with increasing sample size the deviation from Benford take on a higher weight in the calculation. One presupposition for using chi square is that the sample size is large enough; a rule of thumb is that for all digits $np_i^0 \geq 5$, where $p_i^0$ is the expected relative frequency of the Benford distribution. In this case an approximation is only reliable if we have a sample size of over 100. The rejection point for the zero hypothesis that the empirical distribution follows Benford is $\chi^2_{8;0.99} = 20.09$ in the case of $\alpha = 0.01$. Under these conditions many interviewer clusters do not follow the logarithmic distribution. When we look at the GFI values, which are independent of the sample size, we can recognize that small clusters with ≤ 50 digits often have values lower than 0.98. This shows that the approximation is better in bigger samples. Nevertheless, we find that a large proportion of the clusters with true data and two clusters with fakes in samples A and B have values over 0.99. Therefore they all have a good fit to Benford and we can not distinguish between fakes and true clusters.

But indeed, if we compare all fit values of fakes with those of the other clusters, we can recognize that one faked cluster in samples A and B is definitely an outlier: it contains N = 191 digits and with values of $\chi^2 = 84.36$ and a GFI = 0.979, its fit is very bad (the distribution is shown in figure 6). In wave 1 of sample E (1998) three of five clusters with fakes have sample sizes that are too small (≤ 15), but two clusters have sufficient digits (n = 158 and 221) as well as the highest chi-square values of all clusters, with $\chi^2 \geq 50.0$ . In wave 2 of sample E we have only one cluster with fakes but again it has the highest chi-square value of all interviewer clusters. Together with the outlier in sample A, these clusters in sample E are surely candidates for possible fabricated data selected by using a procedure based on Benford's Law.

But we have to interpret this finding with caution. Although the faked clusters have the worst chi-square values, we can't really say that they are "identified". There are several other

[10] This measurement GFI is built in analogy to the well-known goodness of fit measurement GFI for LISREL models. But there the fit value of an actual model refers to a value of the fit function for a model containing only a constant.

clusters close by which are not fabricated but have also bad fit values. Using Benford's Law as fraud detection we make some presumptions:

1. We assume that in principle the first digits of our valid continuous survey data follow Benford's Law.
2. We assume that cheating interviewers don't know this logarithmic distribution or that they are not able to fabricate data, which follow Benford's Law.

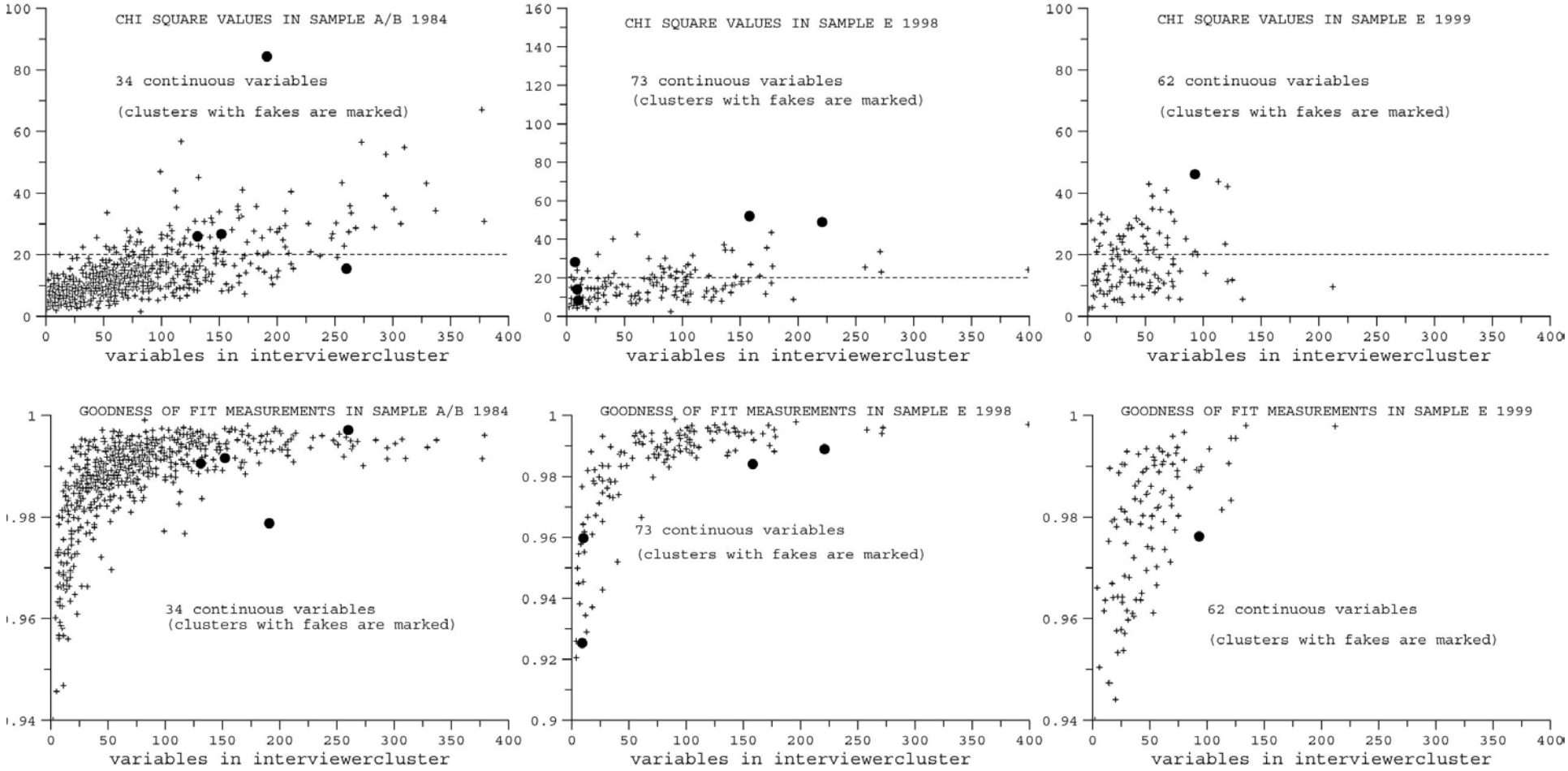


**Figure 7: Chi-square and GFI values for interviewer clusters in Samples A, B and E, individual questionnaire**

The scatterplots show that the major part of the distribution in the interviewer clusters in the observed samples has low chi-square values and that these distributions confirm more or less to Benford's Law. Therefore these empirical findings support assumption 1. Nevertheless a minor part of valid clusters don't confirm Benford's Law and we have to explain this fact. One reason for these misfits may be that some of these valid clusters contain information of respondents in very homogeneous areas. This is possible because the distribution of interviewers and areas are not independent in the SOEP. We can assume that in the case of a homogeneous area the respondents have rather close gross-income values or that the standard deviation of the gross-income variable is lower than in heterogeneous areas. If we take a look at the scatterplot of chi-square values for sample E 1998 in Figure 10 (left hand side) we can observe very close to the two faked clusters a valid cluster (circled) with a chi-square value of 43.47 in the case of n = 177. The distribution of the standard deviation values by the mean of income is shown on the right hand side of Figure 8.[11] We can recognize that the circled cluster has in comparison to the other clusters a very low deviation and that this cluster seems to be rather homogeneously.

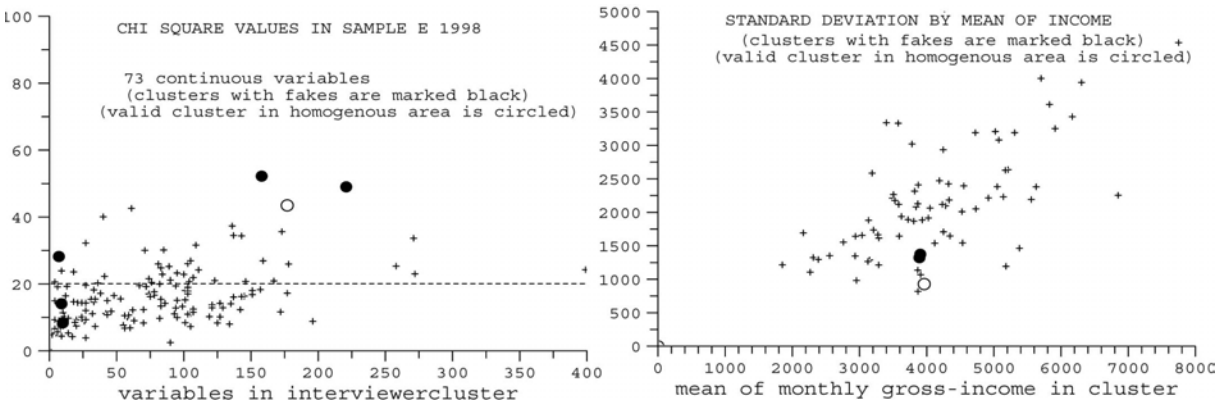


**Figure 8: Chi-square values and standard deviation of gross-income in sample E, individual questionnaire (1998)**

To show the impact of the homogeneity of a cluster on the fit to Benford's distribution we will try to explain the fit values (chi-square values) with the variation coefficient (std/mean), the number of digits of the cluster and a dummy called "fake" which indicates fabricated clusters.[12] Because the faked clusters may affect our results, we also estimate separate regressions without the detected clusters which contain faked data. The estimates for the interviewer clusters in sample A/B and E (1998) are shown in Table 5. We can recognize increasing chi-square values with the number of digits in each cluster and with decreasing variation coefficients. The positive impact of the number of digits is simply caused by the calculation of the chi-square values. The sample size is part of the chi-square formula. But the estimates show that low variation coefficients in homogeneous cluster entail higher chi-square values in the samples with and without the fabricated clusters.

[11] We calculate the standard deviation of the mean values only for cluster with at least three gross-income values. Three small faked clusters have only two gross-income values and they were excluded from the scatterplot.
[12] The variation coefficient is independent of the fake indicator. In sample A/B the mean of the variation coefficient for real data is 0.513 and for the faked cluster 0.492. The difference is not significant (t = 1.092).

**Table 5: Linear regression on the chi-square values for the fit to Benford´s Law in interviewerclusters**

| *Dependent variable: chi-square values* | *Sample A/B* | | | | *Sample E (1998)* | | | |
|---|---|---|---|---|---|---|---|---|
| | *with faked clusters* | | *without faked clusters* | | *with faked clusters* | | *without faked clusters* | |
| | *Coeff.* | *t-value* | *Coeff.* | *t-value* | *Coeff.* | *t-value* | Coeff. | t-value |
| Intercept | 10.210 | 12.51 | 9.971 | 12.83 | 16.485 | 8.23 | 16.433 | 8.18 |
| number of variables (digits) in each cluster | 0.008 | 25.78 | 0.008 | 27.18 | 0.006 | 4.63 | 0.006 | 4.65 |
| variation coefficient for gross-income | -4.523 | -3.27 | -4.090 | -3.09 | -9.330 | -3.03 | -9.342 | -3.02 |
| fake | 15.996 | 4.42 | - | - | 26.889 | 5.15 | - | - |
| $R^2$ (adjusted) | 0.559 | | 0.573 | | 0.387 | | 0.195 | |
| number of clusters | 556 | | 552 | | 107 | | 105 | |

Therefore high chi-square values are not necessary faked clusters. We have to take the homogeneity of the clusters in account.

One explanation of suspicious clusters which passed the quality control measures and which are labelled as non-faked clusters cannot be ruled out: although we believe that the data are not fabricated, this may well be the case. Together with the fieldwork organization which collects the SOEP we will check this in depth. First checks suggest that the suspicious interviewers are not fabricating data, but that they are try to conclude the interviews as quickly as possible. For example one of the suspicious interviewers filled in either data on gross or net monthly labour income. He never filled in both values as expected.

Our preliminary results let us conclude that Benford´s Law might be not an efficient method for detecting faked data, but it might be a new instrument for quality control of the interviewing process.

# 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 (adapted from Schnell 1991 and Kalton 1983). In the case of proportions and means, the overall statistic is the weighted sum of the statistic in the sample with true values and the sample with faked values.

Proportions:

$$P_{sum} = \frac{N_t P_t + N_f P_f}{N} \tag{1}$$

$$P_{sum} = P_t + \frac{N_f}{N}(P_f - P_t)$$

Bias:

$$B_p = P_t - P_{sum}$$

$$B_p = \frac{N_f}{N}(P_t - P_f) \tag{2}$$

$P_{sum}$ total proportion
$P_t$ true proportion
$P_f$ proportion in the faked sample
$N_f/N$ share of fakes on all observations

Equation (2) shows that the possible bias can not 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%. The bias for mean values depends on the proportion of fakes and the measurement scale. If we use satisfaction items with an 11-point scale (0 - 10), for example, in the case of 3% fakes we get a maximum bias of only ± 0.3. Figure 9 shows the proportion of total mean to true mean in terms of the proportion of fakes and means (mean in the faked sample and mean in the true sample). The higher the proportion of fakes and the greater the deviation of means in the faked and true samples, the higher the bias. Nevertheless, the deviation of the factor "mean_sum / mean_true" from one can be only critical if the proportion of fakes is rather high. If the mean in the faked sample is double the mean in the true sample and the proportion of fakes is 0.3, the factor will be 1.6.

Means:

$$M_{sum} = \frac{N_t M_t + N_f M_f}{N} \tag{3}$$

$$M_{sum} = M_t + \frac{N_f}{N}(M_f - M_t)$$

Bias:

$$B_M = \frac{N_f}{N}(M_t - M_f) \tag{4}$$

$M_{sum}$ total mean
$M_t$ true mean
$M_f$ mean in the faked sample
$N_f/N$ share of fakes in all observations

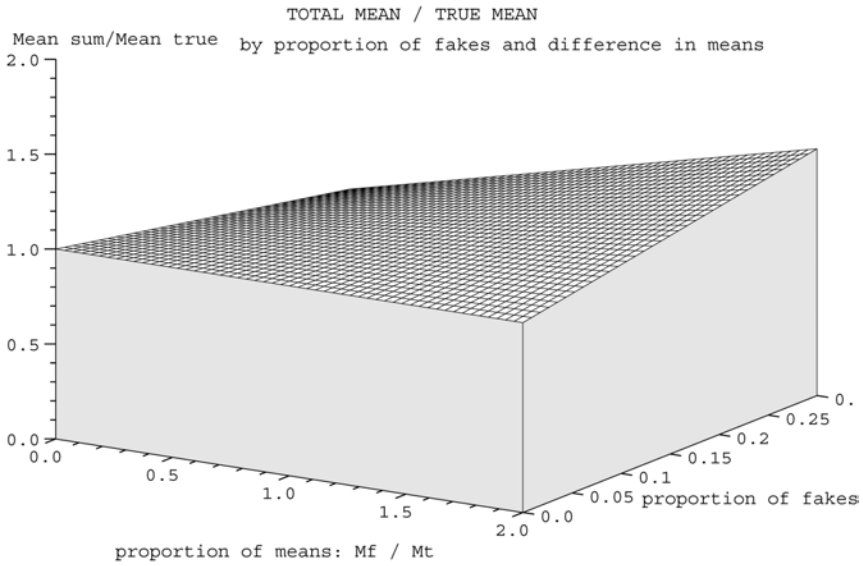


**Figure 9: Total mean / true mean by proportion of fakes and proportion of means**

The total variance is a function of the proportion of faked records in the database, the true and false variance and the squared difference of means in both samples. Figure 10 shows the proportion of the total variance to the true variance. We assume that we have a variable with true mean $M_t = 5.0$ and true variance $\sigma_t^2 = 1.0$, measured on an 11-point scale (e.g. satisfaction items in the SOEP). The false variance is assumed to be 10% lower than the true variance. In the case of small differences in means between the true and faked sample and only a small proportion of fakes, the bias will be negligible. But the figure also shows that at higher proportions of fakes and greater differences in means, we get more seriously biased estimates for the variance.

Variance:

$$\sigma_{sum}^2 = \left(1 - \frac{N_f}{N}\right)\sigma_t^2 + \frac{N_f}{N}\sigma_f^2 + \frac{N_f}{N}\left(1 - \frac{N_f}{N}\right)(M_t - M_f)^2 \tag{5}$$

Bias:

$$B_{\sigma^2} = \frac{N_f}{N}\left(\sigma_t^2 - \sigma_f^2\right) - \frac{N_f}{N}\left(1 - \frac{N_f}{N}\right)(M_t - M_f)^2 \tag{6}$$

| | |
|---|---|
| $\sigma^2_{sum}$ | total variance |
| $\sigma^2_t$ | true variance |
| $\sigma^2_f$ | variance in the faked sample |
| $M_t$ | true mean |
| $M_f$ | mean in the faked sample |
| $N_f/N$ | share of fakes in all observations |

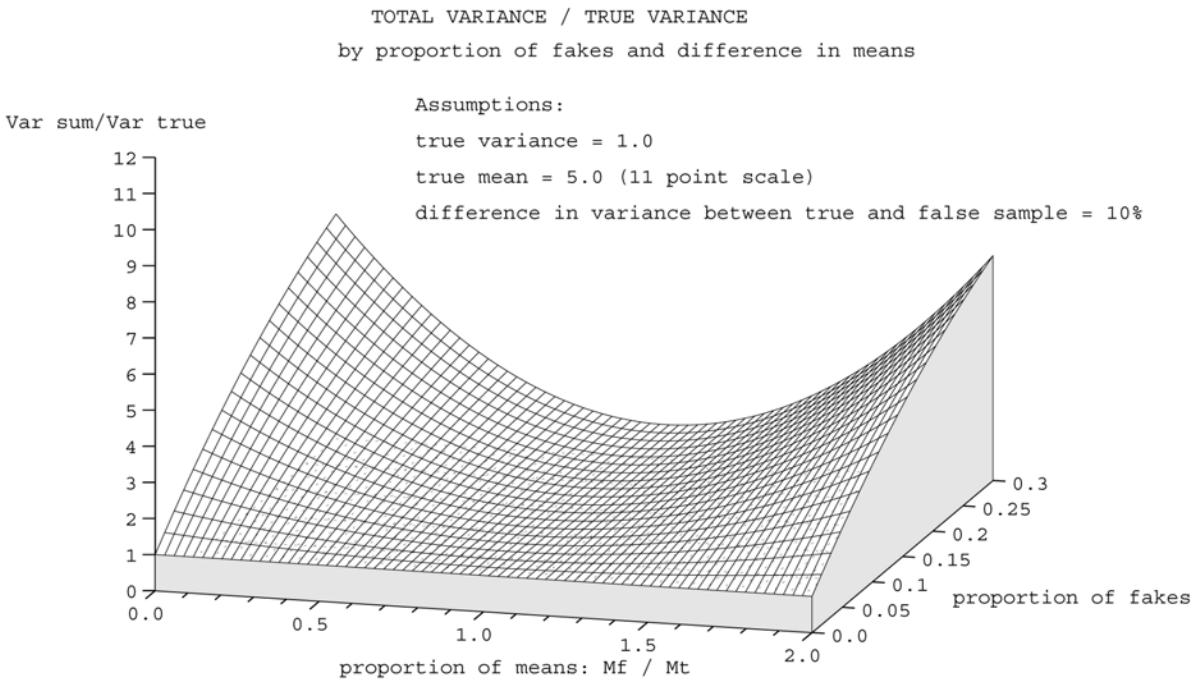


**Figure 10: Total variance / true variance by proportion of fakes and proportion in means**

Next we take a short look at the covariance. The total covariance is a function of the proportion of faked records in the database, the true and false covariance and the product of the difference in means of both variables in the true and faked sample.

Covariance:

$$\sigma_{xy_{sum}} = \left(1 - \frac{N_f}{N}\right)\sigma_{xy_t} + \frac{N_f}{N}\sigma_{xy_f} + \frac{N_f}{N}\left(1 - \frac{N_f}{N}\right)(M_{x_t} - M_{x_f})(M_{y_t} - M_{y_f}) \tag{7}$$

Bias:

$$B_{\sigma_{xy}} = \frac{N_f}{N}(\sigma_{xy_t} - \sigma_{xy_f}) - \frac{N_f}{N}\left(1 - \frac{N_f}{N}\right)(M_{x_t} - M_{x_f})(M_{y_t} - M_{y_f}) \tag{8}$$

$\sigma_{xy\ sum}$ total covariance
$\sigma_{xyt}$ true covariance
$\sigma_{xyf}$ covariance in the faked sample
$M_{xt}$ true mean for variable X
$M_{xf}$ mean in the faked sample for variable X
$M_{yt}$ true mean for variable Y
$M_{yf}$ mean in the faked sample for variable Y
$N_f/N$ share of fakes in all observations

Therefore we have more than three dimensions. But if we assume that the differences in means for both variables are the same, we will get a figure for the total / true covariance quite similar in shape to figure 10.

## 4.2 Empirical bias due to interviewer cheating in SOEP

In this section we present some empirical results using fabricated and true data. We will look first at some descriptive statistics like proportions, means and variances. We analyse only samples A, B and E because the number of fakes in subsample F is too small (N= 8).

### 4.2.1 Proportions

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

Table 6 shows the breakdown of gender responses in fabricated and real samples. The last two columns contain information on the empirical bias and the possible maximal bias[13]. We can detect only a marginal empirical bias. It can be assumed that it is rather easy for cheating interviewers to reproduce responses like respondent's gender because the distribution is known. Hence we will take a look at other variables with more categories. It might be a somewhat more complicated to reproduce the employment status of the SOEP respondents.

**Table 6: Proportion of respondent's gender in fabricated and real samples (individual questionnaire)**

| *Sample A 1984 Respondent's gender* | *Real* | *%* | *Fake* | *%* | *Total* | *%* | *Emp. bias* | *Max. ± bias* |
|---|---|---|---|---|---|---|---|---|
| Male | 4328 | 47.69 | 27 | 45.80 | 4355 | 47.67 | - 0.02 | 0.62 |
| Female | 4748 | 52.31 | 32 | 54.20 | 4780 | 52.33 | 0.02 | 0.66 |
| Total | 9076 | 100.00 | 59 | 100.00 | 9135 | 100.00 | | |

| *Sample B 1984 Respondent's gender* | *Real* | *%* | *Fake* | *%* | *Total* | *%* | *Emp. bias* | *Max. ± bias* |
|---|---|---|---|---|---|---|---|---|
| Male | 1679 | 53.00 | 28 | 62.20 | 1707 | 53.10 | 0.1 | 1.64 |
| Female | 1490 | 47.00 | 17 | 37.80 | 1507 | 46.90 | -0.1 | 1.13 |
| Total | 3169 | 100.00 | 45 | 100.00 | 3214 | 100.00 | | |

| *Sample E. 1998 Respondent's gender* | *Real* | *%* | *Fake* | *%* | *Total* | *%* | *Emp. bias* | *Max. ± bias* |
|---|---|---|---|---|---|---|---|---|
| Male | 932 | 48.80 | 26 | 55.30 | 958 | 49.00 | 0.2 | 2.71 |
| Female | 978 | 51.20 | 21 | 44.70 | 999 | 51.00 | -0.2 | 2.10 |
| Total | 1910 | 100.00 | 47 | 100.00 | 1957 | 100.00 | | |

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

Table 7 shows the distribution of respondent's employment status in samples A and B for fabricated and real data. This variable has seven categories. The highest frequency occurs for the category "full-time employment" with 46.7%, followed by "not employed" with 37.6%. Regular part-time employment responses are only 5.4%, followed by vocational training and unemployed, both with 3.6%. Surprisingly, the distribution for the faked sample is quite similar to the real data. The ranking order of the categories corresponds in both data sets and there are only small deviations in the frequency values, especially for "full-time employment" and "not employed". 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.

**Table 7: Distribution of employment status in Sample A + B, 1984 (fabricated and real data)**

| *Employment status* | *Real* | | *Fake* | | *Total* | |
|---|---|---|---|---|---|---|
| | *N* | *%* | *N* | *%* | *N* | *%* |
| Full-time employment | 5724 | 46.7 | 57 | 54.8 | 5781 | 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 | 439 | 3.6 | 5 | 4.8 | 444 | 3.6 |
| Military, community service | 35 | 0.3 | 0 | 0.0 | 35 | 0.3 |
| Not employed | 4606 | 37.6 | 28 | 26.9 | 4634 | 37.5 |
| Total | 12245 | 100 | 104 | 100 | 12349 | 100 |

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

Table 8 shows the frequency of another item in samples A and B, "the importance of goals in politics". The respondent has to choose between four goals and has to rank these in terms of personal importance. In the real data set we can recognize that the distribution of "peace and quiet" has its highest frequency for the most important goal, "inflation" for the second most important goal, "citizen influence" has equal frequencies for the third and fourth most important goal, and "freedom of speech" has its highest frequency for the fourth most important goal. In the fabricated data we find a rather different distribution, by far the highest frequency for the first goal occurs for "inflation", "peace and quiet" has its highest frequency for the second goal, "freedom of speech" for the third, and "citizen influence" for the fourth. The "importance of goals in politics" is not an objective variable and we can assume that the difference between real and faked data might be caused by personal opinions and philosophies of the cheating interviewers.

Table 8: Frequency of importance of goals in politics[14] – Sample A + B (fabricated and real data)

| | *Peace and quiet* | | *Citizen influence* | | *Inflation* | | *Freedom of speech* | |
|---|---|---|---|---|---|---|---|---|
| | *real* | *fake* | *real* | *fake* | *real* | *fake* | *real* | *fake* |
| Rank 1 | 47.2 | 23.1 | 16.9 | 22.3 | 22.8 | 50.0 | 18.3 | 4.9 |
| Rank 2 | 22.6 | 41.3 | 24.2 | 16.5 | 31.1 | 26.0 | 22.1 | 16.5 |
| Rank 3 | 15.7 | 7.7 | 29.3 | 29.1 | 27.4 | 22.1 | 25.7 | 40.8 |
| Rank 4 | 14.6 | 27.9 | 29.7 | 32.0 | 18.8 | 1,9 | 34.0 | 37.9 |
| | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| N | 11,973 | 104 | 11,865 | 103 | 11,951 | 104 | 11,928 | 103 |

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

### *4.2.2 Means and variances*

Table 9 shows some means and variances in fabricated and real data. We have calculated the means for satisfaction items (11-point scale) and the question about monthly income. The empirical bias in all cases is rather low and negligible; only half of the differences in means between faked and real data are significant.

One finding is consistent: the means for the satisfaction items in sample E are always higher in the fabricated data, and also higher than in the valid data in four of seven cases in samples A and B. With only one exception, the variances in the faked data are lower than in the real data set. These results indicate that cheating interviewers relatively consistently overestimate the satisfaction of respondents. Schnell (1991, 31) argues that the "curbstoners" orientate themselves on stereotypes. He postulates that if this is true, the internal consistence should be higher in the fakes than in the real data. We find for the 11-point likert scale of satisfaction in sample A a value for Cronbach's alpha of 0.672 (sample E: alpha = 0.751), and a value of 0.628 (sample E: alpha = 0.655) in the fabricated data.[15] Hence, in both subsamples the internal consistence is lower in the faked data than in the real data and we can not confirm Schnell's findings in our data.

Table 10 shows the calculated means for the item "importance for satisfaction" (4-point scale). Half of the means based on faked data are higher than in the real data set. Nevertheless, the "fit" of the fabricated data are rather good, the absolute deviation between real and faked mean is only 0.19 on average. An exception is the interviewer's assessment of the importance of "work" for respondent's satisfaction. Here in the fabricated data the mean is 50% higher than in the real data.

[14] The question is: "Even in politics you can't have everything at once. Below are various goals which politics can aim for; if you had to choose between these goals: which seems the most important to you? Which is the second most important? Which is the third most important? And, which is the fourth?"

[15] Cronbach's alpha is a often used coefficient of reliability (or consistency).

**Table 9: Means and variances in fabricated and real data (Sample A, B and E)**

| *Sample satisfaction (11-point scale)* | *A + B 1984* | | | | | | | | *E 1998* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *eal* | | *fake* | | *total* | | *F-test* | *emp.* | *real* | | *fake* | | *total* | | *F-Test* | *emp.* | *max.* |
| | *mean* | $\sigma^2$ | *mean* | $\sigma^2$ | *mean* | $\sigma^2$ | *prob.* | *Bias* | *mean* | $\sigma^2$ | *mean* | $\sigma^2$ | *mean* | $\sigma^2$ | *prob.* | *bias* | *± bias* |
| Health | 7.00 | 7.107 | 8.12 | 3.132 | 7.01 | 7.084 | 0.000*** | 0.01 | 7.00 | 5.681 | 8.06 | 1.974 | 7.02 | 5.617 | 0.002*** | 0.02 | 0,242 |
| Work | 7.65 | 5.208 | 7.94 | 2.171 | 7.65 | 5.178 | 0.279 | 0.00 | 7.17 | 5.772 | 7.56 | 1.949 | 7.18 | 5.681 | 0.411 | 0.01 | 0,242 |
| Household | 6.95 | 5.960 | 6.64 | 8.042 | 6.95 | 5.974 | 0.361 | 0.00 | 6.75 | 5.036 | 6.72 | 2.606 | 6.75 | 4.972 | 0.944 | 0.00 | 0,242 |
| Income | 6.41 | 6.848 | 6.64 | 4.096 | 6.41 | 6.824 | 0.356 | 0.00 | 6.45 | 5.706 | 7.34 | 2.229 | 6.47 | 5.639 | 0.011** | 0.02 | 0,242 |
| Housing | 7.60 | 6.436 | 6.95 | 4.987 | 7.59 | 6.427 | 0.010*** | -0.01 | 7.97 | 3.991 | 8.38 | 0.633 | 7.98 | 3.913 | 0.157 | 0.01 | 0,242 |
| Leisure | 7.29 | 6.497 | 7.05 | 4.590 | 7.29 | 6.481 | 0.340 | 0.00 | 7.40 | 5.011 | 7.87 | 1.809 | 7.41 | 4.938 | 0.148 | 0.01 | 0,242 |
| Products on offer | - | - | - | - | - | - | - | - | 6.44 | 7.963 | 6.55 | 6.122 | 6.44 | 7.915 | 0.781 | 0.00 | 0,242 |
| Public transport | - | - | - | - | - | - | - | - | 6.40 | 8.239 | 7.28 | 4.161 | 6.42 | 8.156 | 0.038** | 0.02 | 0,242 |
| Environmental sit. | - | - | - | - | - | - | - | - | 6.77 | 4.249 | 7.79 | 0.432 | 6.80 | 4.181 | 0.001*** | 0.03 | 0,242 |
| Living standard | - | - | - | - | - | - | - | - | 7.36 | 3.441 | 8.02 | 0.934 | 7.37 | 3.391 | 0.015** | 0.01 | 0,242 |
| Life today | 7.43 | 4.576 | 7.93 | 2.588 | 7.43 | 4.561 | 0.016** | 0.00 | 7.44 | 3.001 | 8.09 | 0.427 | 7.46 | 2.948 | 0.011** | 0.02 | 0,242 |
| Life in 5 years | - | - | - | - | - | - | - | - | 7.23 | 3.836 | 8.06 | 0.539 | 7.25 | 3.772 | 0.004*** | 0.02 | 0,242 |
| *Income* | *real* | $\sigma$ | *fake* | $\sigma$ | *total* | $\sigma$ | *F-Test prob.* | *emp. bias* | *real* | $\sigma$ | *fake* | $\sigma$ | *total* | $\sigma$ | *F-Test prob.* | *emp. bias* | *max. ± bias* |
| Gross income | 2552 | 1569 | 2545 | 1479 | 2552 | 2552 | 0.973 | 0.00 | 4133 | 2574 | 4015 | 1516 | 4129 | 2545 | 0.831 | -4.00 | -138 |
| Net income | 1745 | 11084 | 1760 | 974 | 1745 | 1083 | 0.913 | 0.00 | 2609 | 1549 | 2491 | 910 | 2605 | 1533 | 0.695 | -4.00 | -81 |

Source: SOEP 1984, Sample A,B and 1998, Sample E, individual questionnaire

**Table 10: Means in fabricated and real data (Sample E, individual questionnaire)**

| *Importance for satisfaction (4-point scale)* | *E 1998 means real* | *fake* | *total* | *F-Test prob.* | *emp. bias* | *max. ± bias* |
|---|---|---|---|---|---|---|
| Work | 1.92 | 2.41 | 1.93 | 0.000*** | 0.01 | 0.096 |
| Family | 1.28 | 1.06 | 1.27 | 0.004** | -0.01 | 0.096 |
| Friends | 1.72 | 1.34 | 1.71 | 0.000*** | -0.01 | 0.096 |
| Income | 1.61 | 1.72 | 1.61 | 0.193 | 0.00 | 0.096 |
| Housing | 1.57 | 1.62 | 1.57 | 0.533 | 0.00 | 0.096 |
| Politics | 2.83 | 2.91 | 2.83 | 0.493 | 0.00 | 0.096 |
| Career | 2.26 | 2.51 | 2.26 | 0.090* | 0.00 | 0.096 |
| Leisure | 1.72 | 1.53 | 1.72 | 0.043** | 0.00 | 0.096 |
| Health | 1.15 | 1.00 | 1.14 | 0.007*** | -0.01 | 0.096 |
| Environmentalism | 1.71 | 1.83 | 1.71 | 0.193 | 0.00 | 0.096 |
| Religion | 2.68 | 2.43 | 2.67 | 0.079* | -0.01 | 0.096 |
| Neighborhood | 1.88 | 1.85 | 1.88 | 0.739 | 0.00 | 0.096 |
| Mobility | 1.78 | 1.98 | 1.79 | 0.049** | 0.01 | 0.096 |

### *4.2.3 Covariances and correlations*

In this section we will examine the influence of fabricated data on bivariate statistics such as covariances and correlations. Table 11 shows the covariance and the correlation between net and gross income as well as between gross income and duration of training (years). The relationship between gross and net income is trivial and obvious and apparent in both real and fabricated 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 0.342 in the real data. In the fabricated data only a small negative covariance occurs, or a correlation near zero, respectively. 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.

**Table 11: Covariances and correlations in fabricated and real data of Sample E**

| *Sample E 1998 Gross income* | *Covariance net income real* | *fake* | *Total* | *Correlation net income real* | *fake* | *Total* | *N* |
|---|---|---|---|---|---|---|---|
| Real | 3,390,759 | | | 0.948 | | | 699 |
| Fake | | 1,274,345 | | | 0.924 | | 27 |
| Total | | | 3,310,526 | | | 0.948 | 726 |
| | | | | | | | |
| *Gross income* | *Duration of Training covariance real* | *fake* | *Total* | *(in years) correlation real* | *fake* | *total* | *N* |
| Real | 2181.32 | | | **0.342** | | | 699 |
| Fake | | -38.63 | | | -0.004 | | 27 |
| Total | | | 2110.44 | | | **0.271** | 726 |

Source: SOEP Sample E, 1998, individual questionnaire, true and faked data

### *4.2.4 Linear regressions*

In a further step 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 12 shows the estimated parameters. In the real sample all coefficients have the expected signs and they 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 with adjusted $R^2 = 0.542$. In the fabricated data set (4.7% of the subsample E) we find some differences: the coefficient of duration of training is negative and implausible, the coefficient for working hours is only a third and the coefficient for gender is more than double the coefficient in the true data set. If we leave the fabricated data in sample E we will get biased estimates. In the overall sample the covariates of “age” and “gender” are overestimated and “duration of training” and “working time” are underestimated. The overall fit is lower than in the true data set, the value for adj. $R^2$ declines to 0.378.

**Table 12: Parameters of the linear regression on log gross income (true and fabricated data in Sample E)**

*Sample E - 1998*
*Regression on log gross income*

| | *True* | | *Fake* | | *Total* | |
|---|---|---|---|---|---|---|
| | *coeff.* | *t – value* | *coeff.* | *t - value* | *coeff.* | *t – value* |
| const. | 3.109*** | 12.882 | 4.585** | 2.249 | 4.303*** | 16.687 |
| age | 0.111*** | 10.356 | 0,151 | 1.492 | 0.125*** | 10.230 |
| age squared | -0.001*** | -8.444 | -0.002 | -1.335 | -0.001*** | -8.743 |
| gender (1 - men) | 0.170*** | 3.868 | 0.477** | 2.865 | 0.302*** | 6.267 |
| duration of training (years) | 0.074*** | 9.388 | -0.029 | -0.672 | 0.014* | 1.818 |
| working hours (week) | 0.042*** | 15.341 | 0.0145 | 0.961 | 0.021*** | 8.168 |
| adj. $R^2$ | 0.542 | | 0.296 | | 0.378 | |
| N | 520 | | 26 | | 546 | |

Source: SOEP Sample E, 1998, individual questionnaire, true and faked data

# 5 Summary and conclusions

This paper deals with fabricated 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 a subsample. The share of fabricated data is low in all samples (far less than 1%) and the maximum is 2.4% in sample E. One should note that except for the fakes in sample E, faked data were never disseminated within the widely-used SOEP. The fakes were detected before the data were released. But those fakes are in the original data files – kept at DIW Berlin – and they are a rich source for methodological research.

Because one interviewer was able to fabricate interviews in the first two waves in sample E we were also able to investigate the stability of faked statements. We show that in the case of items about satisfaction and worries the stability coefficient is too low on average and near

zero. We find these low stabilities only in two cases of the data set and conclude that these implausible stabilities can be used to detect fakes.

Furthermore, we applied a new method for discovering frauds. We referred to a recent practice becoming common among accountants to use the Benford distribution of numbers for fraud detection and assign this procedure to survey data. To the best of our knowledge this procedure is used for the first time in combination with survey data to detect frauds. We show that under particular circumstances the first-digit distribution of real data in each interviewer cluster in fact follows the predicted logarithmic distribution. After assigning useful fit values to each interviewer –cluster, we identify one cluster with fakes in sample A and two clusters in sample E as outliers with worst fit values. These clusters are candidates for possible "fabricated clusters", identified by Benford's Law. In addition we show that bad fit values are not necessarily fakes, because we have to take the homogeneity of each cluster and the number of digits in each cluster into account. Our preliminary results let us conclude that Benford´s Law might be not an efficient method for detecting faked data, but it might be a new instrument for quality control of the interviewing process.

Finally, we analyse the impact of faked interviews on survey results. We show that the impact of interviewer cheating on proportions can not be greater than the proportion of the fakes in the sample. Overall we could observe that the bias for proportions is very small and negligible in the SOEP, not only because the share of fakes is low, but because the "quality" of fakes is high. Interviewers who cheat often have an idea of the distribution of a particular variable such as "employment status" and can successfully reproduce the frequencies of this variable in the data they deliver to the fieldwork organization. The bias of mean values depends on the proportion of fakes and the measurement scale. For the satisfaction items we can show that cheating interviewers consistently overestimate respondents' satisfaction and that the mean in most cases is higher than in the true data set.

Whereas the bias of proportion and means are 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 could show that some cheating interviewers are swamped with multivariate statistics and failed 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 consequent parameters can be seriously biased. Therefore we find empirical evidence for the finding by Schnell (1991), based on his simulation results, that even small proportion of fake interviews are an important problem in multivariate survey statistics.

# 6 References


Benford, Frank 1938: The Law of Anomalous Numbers, Proceedings of the American Philosophical Society, 78:4, 551-572.

Biemer, Paul P. and S. Lynne Stokes, 1989: The Optimal design of Quality Control Samples to Detect Interviewer Cheating. Journal of Official Statistics, 5:1, 23-39.

Bushery, John M., Jennifer W. Reichert, Keith A. Albright, and John C. Rossiter, 1999: Using Date and Time Stamps to Detect Interviewer Falsification. Proceedings of the American Statistical Association (Survey Research Methods Section), 316-320.

Cantwell, Patrick J., John M. Bushery, and Paul P. Biemer, 1992: Toward a Quality Improvement System for Field Interviewing: Putting Content Reinterview Into Perspective. Proceedings of the American Statistical Association (Survey Research Methods Section), 74-83.

Crespi, L.P. 1945: The Cheater Problem in Polling. Public Opinion Quarterly, Winter: 431-445.

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

Evans, Franklin B. 1961: On Interviewer Cheating. Public Opinion Quarterly, 25: 126-127.

Hill, Theodore P. 1995: A Statistical Derivation of the Significant-Digit Law, Statistical Science, 10: 354-363.

Hill, Theodore P. 1996: The First –Digit Phenomenon, American Scientist, 86: 358-363.

Hill, Theodore P. 1999: The Difficulty of Faking Data, Chance, 26: 8-13.

Hood, Catherine C. and John M. Bushery, 1997: Getting more Bang from the Reinterview Buck: Identifying "At Risk" Interviewers. Proceedings of the American Statistical Association (Survey Research Methods Section), 820-824.

Kalton, G. 1983: Compensating for Missing Survey Data. Ann Arbor: Institute for Social Research.

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

Moore, Jeffrey C. and K.H. Marquis, 1996: The SIPP Cognitive Research Evaluation Experiment: Basic Results and Documentation. Working-Paper No. 212, U.S. Department of Commerce, Bureau of the Census.

Newcomb, Simon 1881: Note on the Frequency of Use of the Different Digits in Natural Numbers, American Journal of Mathematics, 4: 39-40.

Nigrini, Mark 1999: I've got your number, Journal of Accountancy, 187: 79-83.

Pinkham, Roger 1961: On the distribution of the first significant digits, The Annals of Mathematical Statistics, 32: 1223-1230.

Reuband, Karl-Heinz 1990: Interviews, die keine sind - "Erfolge" und "Mißerfolge" beim Fälschen von Interviews. KZfSS, 4: 706-733.

Schnell, Rainer 1991: Der Einfluss gefälschter Interviews auf Survey Ergebnisse, ZfS, 20:1, 25-35.

Schräpler, Jörg-P. 1999: Das Befragtenverhalten im Sozio-oekonomischen Panel. Dissertation. Ruhr-University Bochum.

Schräpler, Jörg-P. and Gert G. Wagner 2001: Das Verhalten von Interviewern – Darstellung und ausgewählte Analysen am Beispiel des Interviewer-Panels des Sozio-oekonomischen Panels (SOEP). Allgemeines Statistisches Archiv, 85, (1), 45-66.

Schreiner, Irwin, Karen Pennie, Jennifer Newbrough, 1988: Interviewer falsification in Census Bureau surveys. Proceedings of the American Statistical Association (Survey Research Methods Section), 491-496.

Stokes, S. Lynne, Patty Jones, 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.

Turner, Charles F., James N. Gribble, Alia A. Al-Tayyib, and James R. Chromy, 2002: Falsification in Epidemiologic Surveys: Detection and Remediation (Prepublication Draft). Technical Papers on Health and Behavior Measurement, No. 53. Washington DC: Research Triangle Institute.