Do Falsifiers Leave Traces? Finding Recognizable Response Patterns in Interviewer Falsifications

Walzenbach, Sandra

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Walzenbach, S. (2021). Do Falsifiers Leave Traces? Finding Recognizable Response Patterns in Interviewer Falsifications. *Methods, data, analyses : a journal for quantitative methods and survey methodology (mda)*, *15*(2), 125-160. https://doi.org/10.12758/mda.2021.02

methods, data, analyses | 2021, Vol. 15(2), pp. 125-160 DOI: 10.12758/mda.2021.02

Do Falsifiers Leave Traces? Finding Recognizable Response Patterns in Interviewer Falsifications

Sandra Walzenbach
University of Konstanz

Abstract

Fraud by interviewers is a ubiquitous threat to data quality in survey practice, whenever face-to-face surveys are conducted. Particularly if interviewers use stereotypes about respondents to fill in questionnaires, falsifications can limit the variety of possible answers, lead erroneously to significant correlations and distort survey results.

In addition to external control mechanisms to detect fraud (such as postcards or time stamps) more recent research has started to also consider internal indicators (such as the number of missing values or open answers) as a monitoring strategy. This latter approach relies on ex-post statistical analyses and implicitly assumes that falsifiers apply rational behavioral strategies which result in detectable response patterns. This study examines to what extent fieldwork monitoring can benefit from such approaches, by empirically assessing how effective different indicators are at detecting known cases of fabrication.

In contrast to most previous research, which often relies on laboratory fabrications, this study uses authentic cases of detected interviewer fraud from a survey on the fairness of earnings conducted in Germany. The main goal of this study is to examine to what extent the falsifiers’ attempts to produce unsuspicious data led to recognizable response patterns. For this purpose, we test a wide range of indicators that could potentially identify falsifications: avoidance of extreme categories and open text-based answers, low rates of item-nonresponse, strategic use of filter questions to shorten the questionnaire and non-compliance of responses to numeric questions with Benford‘s Law. Furthermore, we compare authentic and fabricated interviews according to their values on a social desirability scale and report results from an innovative trick question that was especially designed to detect falsifiers.

Keywords: interviewer falsification, interviewer fraud, interviewer effects, response patterns, statistical methods, data quality

Interviewer Falsifications as an Omnipresent but Seldomly Discussed Topic

Whenever interviewers are employed to collect survey data, researchers have to face the problem that their interviewers do not necessarily follow the same interests as they do. In fact, the interviewers‘ goals might even contradict the researchers‘ in many aspects (Winker 2016). While researchers are interested in unbiased data, ideally obtained from random samples with high response rates, interviewers might, in the worst case, aim to waste as little time as possible to conduct the necessary amount of interviews. Delivering quality data might not be their primary concern. In his book on 19 years of professional experience as an interviewer, Dorroch reports how interviewers complete their daily tasks with the least investment of effort and time possible (Dorroch 1994). Apart from being a nightmare for every researcher, his narration clearly points out that, from an interviewer’s perspective, deviating from the interviewer guidelines and falsifying data can be a very beneficial rational decision.

Estimates of the actual amount of fabricated data in surveys vary to some extent, but most authors assume a share of between less than one and seven percent (Finn & Ranchhod 2017; Koch 1995; Schräpler & Wagner 2003; Schnell 1991; Schreiner, Pennie, & Newbrough 1988), while some mention possible numbers of above 50 percent depending on the survey and its supervision capacities (De Haas & Winker 2016).

Although data falsification is an omnipresent problem in survey research and sometimes even receives some attention in popular media (as in a prominent feature by the German Spiegel magazine in February 2018), little scientific literature has focused on the topic. This is particularly surprising because previous research has shown that fabricated data can systematically bias survey results: Schnell (1991) quantifies the potential threat of falsifications for data quality by varying the share

Acknowledgements

A first version of this paper was written in Katrin Auspurg’s seminar on “Survey Research Methods” at the University of Konstanz. I would like to express my gratitude for her support and guidance. The survey on the fairness of earnings was part of the project “The factorial survey as a method for measuring attitudes in population surveys”, which received funding from the German Research Foundation (DFG). The survey was designed, organised and monitored by Carsten Sauer, Katrin Auspurg, Thomas Hinz, and Stefan Liebig. Thomas Hinz had the crucial idea for the trick question that is discussed on page 13.

Direct correspondence to

Sandra Walzenbach, University of Konstanz, Department of Sociology,
Universitätsstraße 10, 78464 Konstanz, Germany
E-mail: sandra.walzenbach@uni-konstanz.de

of (laboratory) fabrications in a data set. He concludes that a share of five percent hardly affects univariate analyses. Multivariate analyses, however, were much more susceptible to bias (also see Reuband 1990). This paper therefore aims to contribute to a debate that we consider necessary in order to find an adequate way of dealing with falsification by interviewers.

Survey agencies and researchers usually apply a variety of monitoring strategies to deal with potential interviewer fraud (AAPOR 2003; Murphy, Biemer, Stringer, Thissen, Day, & Hsieh 2016). For our purposes, it is sufficient to distinguish between what we will call external and internal control mechanisms:

- *External control mechanisms* are external to the substantive answers in the questionnaire. Widely used techniques include recontacting respondents via postcard or phone call (e.g. Koch 1995: 91f), the storage of paradata such as time stamps to determine the length of an interview (e.g. Hood & Bushery 1997: 820f) or the number of conducted interviews per day (Bushery, Reichert, Albright, & Rossiter 1999: 317f) and the time gap between them. Some authors suggest that more experienced interviewers might use more sophisticated forms of falsifications (Schreiner et al. 1988), a consideration that leads Hood & Bushery (1997) to keep track of suspiciously high numbers of ineligible households that the interviewer might have misclassified to avoid hard-to-reach respondents. Less common but promising observational methods such as audio recordings or GPS tracking also belong to this category (Thissen 2014; Thissen & Myers 2016; Wagner, Olson, & Edgar 2017).
- *Internal control mechanisms* refer to statistical ex-post analyses of the substantive answers from the questionnaire. In contrast to external control mechanisms, the analysis of internal response patterns aims to develop a technique that can identify falsifications merely on the basis of the completed questionnaires themselves (e.g. Bredl, Winker, & Kötschau 2012; De Haas & Winker 2016; Kosyakova, Olbrich, Sakshaug, & Schwanhäuser 2019). This more controversial approach draws on rational choice theory and the assumption that the falsifier's attempt to produce unobtrusive and unsuspicious data results in certain recognizable response patterns.

Internal control mechanisms are not meant to replace external checks. Rather, they are a cost-efficient supplement to external control mechanisms that can be useful to preselect suspicious interviewers for further, more targeted examination.

Research Objective and Approach: Do Falsifiers Leave Traces?

So far there is no scientific consensus on a superior monitoring strategy to deal with interviewer fraud (Murphy et al. 2016). Instead there is a variety of coexisting measures that either aim to prevent or retrospectively discover falsifications, particularly when it comes to internal control mechanisms, that is, *indicators* that are internal to the collected questionnaire data. Some authors have suggested principal component analysis to examine similar response patterns on ordinal response scales (Blasius & Thiessen 2013) or cluster analyses that combine several statistical indicators to identify interviewers at risk (e.g. Bredl et al. 2012). However, these latter approaches often suffer from a high number of false positives, particularly in settings where the individual interviewers complete few interviews and the share of fraudulent interviewers is low (De Haas & Winker 2014; Storfinger & Winker 2013) and no a priori restriction on the number of falsifiers is defined (De Haas & Winker 2016). In addition, it is unclear which indicators are best suited for cluster analyses (see Menold, Winker, Storfinger, & Kemper 2013 for a simulation study testing different combinations of indicators on a laboratory sample with 50% falsifications).

As will be argued in more detail in the following section, the little research that empirically tests such indicators has produced somewhat contradictory results. This paper therefore focuses on statistical ex-post analyses of response patterns and examines to what extent the response patterns in fabricated data reflect the typically assumed 'rational' interviewer behavior that translates into detectable peculiarities: Do falsifiers leave traces that make them identifiable?

We will empirically test five internal indicators using a survey on the fairness of earnings in Germany. These data are particularly suitable for the present research, because they contain (at least) 44 authentic cases of interviewer fraud, which can be tested for typical response patterns. Apart from more conventional external control mechanisms, such as control postcards and time stamps (by which these falsifications were discovered), the questionnaire also contained a trick question on the income inequality in Europe. This rather innovative attempt to identify falsifiers merely by their response patterns will be discussed in more detail later on.

A major limitation of most previous studies on interviewer falsifications is that they rely on artificial laboratory experiments, in which arbitrarily chosen respondents (often university students) are asked to fabricate data (for a recent exception see Schwanhäuser, Sakshaug, Kosyakova, & Kreuter 2020). It is a crucial advantage that the data at hand allow us to analyze data from authentic interviewers with intentions to fake data in a real life situation.

Previous Research

Types of Interviewer Fraud

Generally, interviewer fraud is defined as an intentional deviation from the interviewer guidelines (AAPOR 2003; Gwartney 2013). These deviations can occur at different steps of the interview process and vary in their degree of severity (for an extensive list see Murphy et al. 2016). The AAPOR (2003) talks about "a continuum of severity of falsification" (page 2). For the purpose of this study, milder forms of interviewer deviations such as rephrasing questions, failing to record verbatims or allowing refusal and item-nonresponse will not be discussed in more detail. Although focus group interviews among interviewers suggest that such minor deviations are the most common type of interviewer falsification (Nelson & Kiecker 1996), it can be hard to determine in an individual case if e.g. by rephrasing, an interviewer intended to help a respondent or to falsify data.

Leaving minor interviewer deviations aside, Schnell (1991) distinguishes between three essential types of falsifications, into which most other classifications (e.g. AAPOR 2003; Schreiner et al. 1988) can be condensed:

- *complete falsifications*, meaning that the interviewer fills in the whole questionnaire without contacting the designated respondent
- *partial falsifications*, for which the interviewer collects some crucial information, either directly from the respondent or from someone who knows him/her, in order to complete the remaining questions alone
- cases in which the *random procedure to select respondents is ignored*, meaning that instead of the target subject someone else is interviewed or the eligibility of a potential hard-to-reach respondent is misreported

The three types of interviewer fraud differ in two aspects: In how demanding it is (for the interviewer) to produce them and in how demanding it is (for the researcher) to detect them (Schnell 1991: 27-29). This last distinction is crucial for this paper insofar as falsifications that ignore the random selection process are impossible to discover solely by means of statistical ex-post analyses of the data. Real respondents will produce unsuspicious response patterns no matter if they were part of the random sample or not. Hints for this specific kind of fraud can only be gained with the aid of external control mechanisms. As a consequence, the approach presented here will only be helpful for the identification of certain types of falsifications. At the same time, the analyses of response patterns cannot be - and do not intend to be - a replacement for external control mechanisms but a first step to identify suspicious cases. Moreover, partial falsifications will be harder to identify than complete falsifications. For partial falsifications, detection will be easier to accomplish the larger the falsified fraction of questions within an interview (De Haas & Winker 2014).

Interviewer Fraud as a Rational Behavior

Fraud can be considered the result of a rational decision process, in which the interviewers react to the situational circumstances they encounter. According to subjective expected utility theory (Kroneberg & Kalter 2012; Esser 1999: 247-275), a wide version of rational choice theory, such a decision process is a function of the following factors:

- the alternative actions to choose from
- the subjective utilities associated with these alternatives
- the costs associated with each alternative
- the perceived probability that an action can actually be carried out and thus leads to the expected utility

The worst case scenario from a researcher's perspective would be an interviewer who aims to complete the job in as little time and with as little effort as possible. In line with theoretical assumptions, the respective survey literature has identified a variety of circumstances that might make fraud more likely. Gwartney (2013) argues that "calculating cynics" (page 203) who fake frequently and systematically are rare in practice. She believes that most interviewers rather fake occasionally, when their ethics break down in difficult situations. However, she acknowledges that the interviewers' working environment can strongly encourage them to deviate from instructions. Already decades ago, Crespi (1945) argued that researchers and survey agencies can change the "demoralising" circumstances, in which interviewers make their decisions. Similarly, Koch pointed out in the 1990s that it would be wrong to solely blame the interviewers. He considered defective interviewer training, long and poorly designed questionnaires, and meagre salaries a part of the problem (Koch 1995: 102). Gwartney (2013) adds complicated sampling procedures and software, performance and deadline pressures, and a lack of appreciation and support from fieldwork agencies to the list. Interesting empirical evidence for these problems is provided by a qualitative study among interviewers (Nelson & Kiecker 1996) and a field experiment that manipulates the interviewers' working conditions (Menold, Landrock, Winker, Pellner, & Kemper 2018). In addition, some recent work highlights the importance of work ethics and moral values which should be articulated by researchers and supervisory staff (AAPOR 2003; Gwartney 2013; Murphy et al. 2016).

Empirical Evidence on Suspicious Response Patterns

It has been suggested that interviewers do not only act rationally when they make the decision to (or not to) falsify, but also while they are trying to produce unsuspicious data: "Interviewers who falsify will try to keep it simple and fabricate a minimum of falsified data" (Hood & Bushery 1997: 820). If this is the case, faked data would show statistically detectable differences to properly completed questionnaires. The underlying question is *how* interviewers fabricate, that is, which typical response patterns make them identifiable by statistical analyses. When discussing the issue, researchers commonly refer to the same response patterns in line with the assumption of a rationally acting falsifier. The empirical evidence on these response patterns, however, is far from conclusive. Results from different studies are often inconsistent and sometimes clearly contradictory.

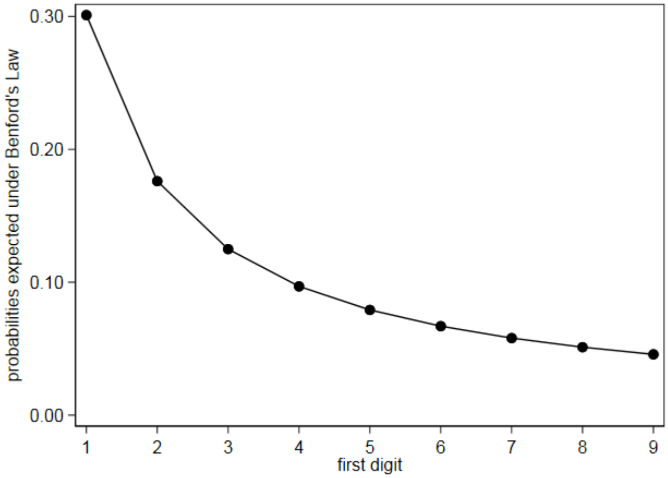
The response patterns that are typically expected from falsifiers will now be discussed in more detail, referring to previous studies that tested or used such patterns as indicators for fraud. The aim of the following section is twofold: It will introduce the internal indicators that will be empirically tested later on and it gives a first rough impression about how the indicators performed in fraud detection in the literature to date.

Nonconformity with Benford's Law (1)

Benford's Law implies that multi-digit numbers are more likely to start with a small than with a high digit (Benford 1938, Diekmann & Jann 2010). The frequency that *a* is the first digit in multi-digit numbers follows the distribution:

$$F_a = \log_{10}\left(\frac{a+1}{a}\right)$$

As Graph 1 shows, the first digit is one in approximately 30% of all numbers. With each next higher digit, the values continue to decrease.



Graph 1 Frequency of first place digits according to Benford's Law

What makes Benford's Law a potentially helpful tool for fraud detection is that it is perceived as "quite counterintuitive" (Nigrini 1999): Falsifiers should expect equal frequencies for all possible first digits in a number. Based on this assumption, Benford's Law has not only been used for fraud detection in survey data but also in regression coefficients of journal publications (Bauer & Gross 2011; Diekmann 2007). The extent to which values comply with Benford's Law is then assessed by a chi-squared test, measuring the difference between the observed and the expected distribution.

Various studies examine monetary values to identify falsifiers in the German Socio-Economic Panel (GSOEP): While Schäfer, Schräpler, Müller, & Wagner (2005) and Schräpler (2011) are mildly positive about their results, Schräpler & Wagner (2003) describe the method as "not efficient". Interestingly, they extend the analysis to the second digit distribution, which does not bring about convincing results. Also the true data contains far too many zeros due to rounding errors. A similar problem leads Porras and English (2004) to successfully test different variants of Benford's Law, including one that excludes the digit 5. Bredl et al. (2012) come to the conclusion that their falsifiers diverge more from Benford's Law than real respondents, although the difference is not very big. All in all, empirical evidence is mixed. Nonetheless, Benford's law is often simply assumed in scientific studies and used as an adequate means of detecting fraud, without verifying that, on the one hand, the real data follow the distribution and, on the other hand, that the falsifications do not (Diekmann & Jann 2010: 398).

Avoiding extreme categories (2)

It is usually assumed that falsifiers avoid the more obtrusive extreme categories in ordinal response scales, leading to smaller variances in faked data. Most empirical studies seem to find such differences in either the number of extreme categories or variances (e.g. Bredl et al. 2012; Kemper & Menold 2014; Schäfer et al. 2005). However, there are some exceptions: Menold & Kemper (2014) report inconsistent results and Schnell (1991) cannot find any effect, although he theoretically argues that falsifiers should generally underestimate the heterogeneity of respondents.

Strategic use of filter questions to shorten questionnaire (3)

An empirically rather uncontested hypothesis is that falsifiers make use of their knowledge about the filter branches in a questionnaire to find the shortest and easiest way through it (Menold & Kemper 2014; Brüderl, Huyer-May, & Schmiedeberg 2013; Josten & Trappmann 2016). In other words, it should be a promising strategy to take a closer look at the answers that interviewers gave to "gate questions" (Weinauer 2019), that is, questions that result in a list of follow-up questions. Alternatively, the number of inapplicable questions can be examined.

Avoiding item-nonresponse (4)

For the remaining questions that cannot be avoided by filter questions, contrastingly, it is believed that falsifiers provide answers more consistently throughout the questionnaire than real respondents because they do not want to raise suspicion. They hence should produce less item-nonresponse. Bredl et al. (2012) confirm this empirically, while Schnell (1991) finds the opposite effect.

Avoiding open answers (5)

Another common idea is that falsifiers should avoid open answers. This makes sense from a rational choice perspective: On the one hand because fictitious answers might be easily detected and on the other hand because it is comparatively burdensome and time-consuming to invent and write down a plausible answer. Empirically, things seem less clear: While Bredl et al. (2012) find fewer "other"-answers in their falsifications, Menold & Kemper (2014) come to opposite conclusions.

Hypotheses

After discussing these most common internal indicators, in the light of rational choice theory and with respect to prior findings from previous research, the subsequent part of this paper will subject the indicators to an empirical test. If the traditional assumptions of rational behavior hold, we generally expect to see the following response patterns in falsified data:
Compared to authentic respondents, …

1) … falsifiers violate Benford's Law when they report numbers.

2) … falsifiers choose fewer extreme categories on ordinal response scales.

3) … falsifiers use filter branches to shorten the questionnaire.

4) … falsifiers produce less item-nonresponse within their filter path.

5) … falsifiers give fewer open answers.

Trick question (6)

Apart from these rather commonly used indicators, we will also empirically test an unconventional approach to fraud detection, namely, a trick question which was deliberately designed and implemented as a potential method to identify fraud in this specific survey (for a different approach to trick questions where respondents are asked about fictitious words or newspapers, see Ziegler, Kemper, & Rammstedt 2013; or Menold & Kemper 2014; Winker, Kruse, Menold, & Landrock 2015 for implementations).

Respondents were asked for the European country with the highest income inequality. However, on their training, interviewers received false information about likely responses. As a consequence, we would expect falsifiers to avoid the presumably rare true answer “Portugal”, while unsuspicious Eastern European countries should be mentioned more often than in real interviews.
These considerations result in the following additional hypotheses:

6a) … falsifiers avoid the presumably rare true answer “Portugal”.

6b) … falsifiers overestimate the share of mentioned Eastern European countries.

More details on the trick question are provided in the section on data and methods.

Data and Methods

To test the potential of the discussed indicators for fraud detection, we use data from a cross-sectional survey on the fairness of earnings, in which authentic cases of fraud were detected during fieldwork. The survey was part of the project “The factorial survey as a method for measuring attitudes in population surveys”, funded by the German Research Foundation (DFG). The questionnaire contained single-item and vignette questions on income-related fairness perceptions, some knowledge questions about income and labour in Germany, information on the respondent’s own income, occupation and working environment, as well as questions on the respondent’s socio-demographic background and a social desirability scale. All in all, the questionnaire was of moderate length: 70% of face-to-face respondents completed the questionnaire in 20 to 30 minutes.

The survey was conducted nationwide among the residential population of Germany aged 18+. In about 50% of cases, data were collected by interviewers in computer-assisted face-to-face interviews (CAPI). The sampling strategy comprised the random selection of 129 sample points throughout Germany, a random route procedure and a Kish–selection grid. The other half of respondents was recruited via telephone by means of random digit dialing in combination with a Kish-selection grid. This group completed a self-administered paper or online questionnaire (PAPI/CAWI). Since there was no possibility for interviewer falsification in the self-administered sample, the present project focuses on the 821 interviews conducted in the face-to-face setting. The 803 self-administered questionnaires are only occasionally mentioned for purposes of comparison.

The fieldwork monitoring comprised a range of external control mechanisms: In a first step, re-contact via postcard and paradata of interview time and duration were used to identify suspicious cases. In a second step, the suspicious cases were subjected to repeated contact attempts by telephone and follow-up checks of the random route. This process identified 44 falsifications. These falsifications were admitted by the interviewers and consequentially deleted in consultation with the survey agency.

Fraud occurred in ten different sample points, seven of which were completely and three partially removed from the official data set. Since interviewer characteristics were not made available for the faked data, we need to make the (reasonable) assumption that each sample point was assigned to one interviewer in order to correct for clustered standard errors in the significance tests throughout the empirical analyses of this paper. If this assumption is true, the faked data were produced by ten different falsifiers. Out of these, seven interviewers falsified all of their work (five to seven interviews), while three only falsified one or two of their interviews.[1] To capture this pattern adequately and allow for the fact that interviewers have not necessarily faked all of their assigned interviews, the subsequent analyses will be carried out on the interview level. Looking at the existing literature on interviewer falsifications, this is a somewhat unusual approach. However, it best reflects the nature of our data and accounts for the fact that the number of conducted interviews per interviewer is too small to run reliable analyses at the interviewer level.

With regard to the explanatory variables, indicators are generally obtained by summing up over all available questions and separately for each interview. Put concretely, this is done for the first and second digits of the monetary values, the extreme categories on ordinal scales, inapplicable questions, missing values in mandatory questions, and the prevalence of open answers. We will use all of these internal indicators to compare real and fake interviews in order to obtain information on the extent to which falsifiers leave detectable traces within questionnaires. First, this will be done by descriptive comparisons between the two groups (page 13-20). The statistical tests reported for the indicators 2 to 6 are tests for mean comparisons, Somers'D or tests of proportions dependent on the variable's level of measurement. All of them use cluster-robust standard errors to account for the fact that interviews are nested within interviewers. In a second step, multivariate analyses are presented. In a logistic regression model, in which the dependent variable indicates whether an interview was actually conducted or fabricated, we will assess the relative importance of the explanatory variables and identify the most promising indicators for fraud detection. Again, cluster-robust standard errors are applied to account for the fact that observations are not independent but nested within interviewers.

[1] Apart from one exception, falsifiers had a workload of five to seven interviews per interviewer. There is one falsifier with a workload of 13 interviews who falsified one of them. On average, unsuspicious interviewers completed more, that is, 12 interviews. For unsuspicious interviewers, numbers ranged from two to 28 interviews per interviewer.

Before moving on to the results, the remaining part of this subsection will provide more details on the concrete survey questions that the individual explanatory indicators rely on. The full question wording is provided in Appendix B.

- **Nonconformity with Benford's Law (1)**
  The questionnaire contained five monetary variables that can be checked for their compliance with Benford's Law: The estimated average monthly gross income for a full position in Germany, the respondent's own monthly gross income, the own gross income that the respondent would perceive as fair, the net household income per month and the household income necessary to pay for recurring expenses. For all of these items, respondents were asked to provide an open answer. While the available number of numeric values is too small to conduct analyses at the level of individual interviews or interviewers, comparing the real and the faked data as a whole can produce valuable insights concerning the usability of the Benford distribution for fraud detection.
- **Avoiding extreme categories (2)**
  To examine the shares of extreme response categories, we use two item batteries with 7-point Likert response scales. These questions were answered by all respondents, that is, they could not be inapplicable. In the first item battery, respondents were asked to evaluate to which extent certain person characteristics should have an impact on an employee's fair gross income. This task was completed for eleven different characteristics (e.g. sex and education) using an ordinal scale ranging from 0 ("not important at all") to 6 ("very important"). This item battery is followed by a social desirability scale, in which the respondents assessed their own personality on the basis of six statements that they evaluated on a scale from 0 ("not applicable at all") to 6 ("fully applicable"). Graph 3 shows the frequency of each response category for real and fabricated interviews across all 17 items.
- **Strategic use of filter questions to shorten questionnaire (3)**
  Twenty of the survey questions can in principle be skipped due to filters. A falsifier who knows the questionnaire could answer the filter questions strategically to shorten the questionnaire, resulting in higher numbers of inapplicable questions within filter paths.
- **Avoiding item-nonresponse (4)**
  There were 43 closed-ended questions that were mandatory for every respondent and could be subjected to an examination of item-nonresponse. The respective indicator consists of the sum of missing values in these questions.
- **Avoiding open answers (5)**
  There are seven open answers in the questionnaire that can be used to test hypothesis 5. Three of them are open questions in the stricter sense of the word, namely the job title and description of the respondent's current (or last) occupation, a feedback question at the end of the questionnaire and a trick ques-

tion on the European country with the highest income inequality (which we will come back to in in the results section). Apart from that, there were seven occasions in which respondents could specify additional options ("other, please specify:_________________"). This was possible for their current occupation, their main place of residence since birth, the sources of their household income, their party preference, their partner's occupation, their highest educational degree, and their vocational qualifications. However, the latter three were not used by any respondent. We are hence left with seven potential open answers to analyze. For the descriptive analyses, the overall number of open answers serves as an indicator. For the regression model, we distinguish between completely open questions and text fields, where respondents could specify "other" options.

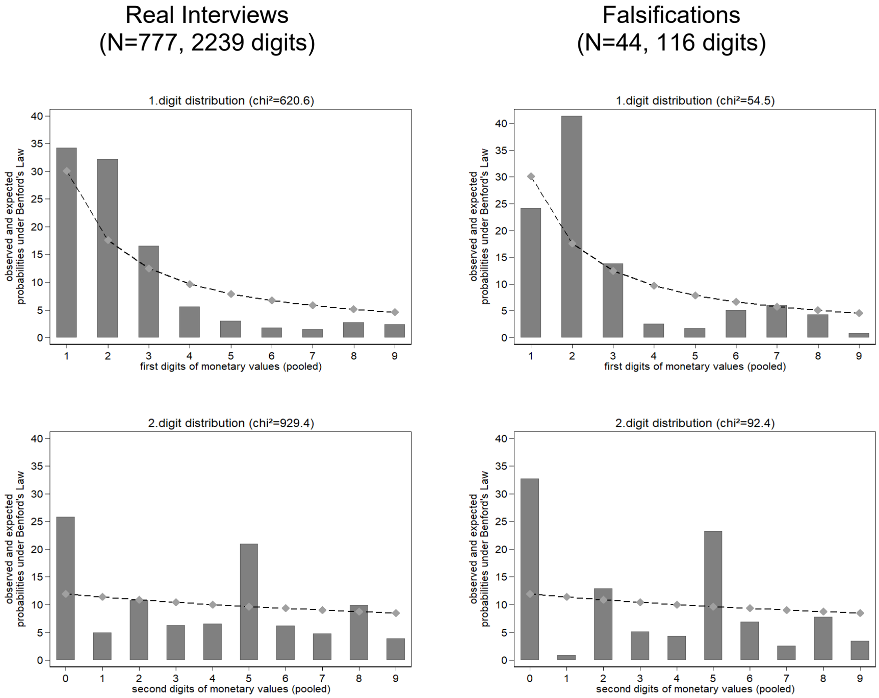
- **Trick question (6)**
  Implementing a trick question was an innovative attempt to detect fraudulent data in the setting of this specific survey. As part of an item battery that measured the respondents' general knowledge of the survey topic, they were asked for the European country with the highest income inequality. In the training that all the interviewers underwent before the fieldwork period, interviewers were informed about the right answer (which was Portugal at the time). However, they were also told that only experts know this and that respondents would usually guess an „Eastern European country such as Poland or Romania" (Sauer, Auspurg, Hinz, & Liebig 2010: 79). We will use the answer to the trick question to examine if such a trick question is helpful to detect fraud. This would be the case if falsifiers were more likely then real respondents to provide a presumably unsuspicious country, while avoiding the true but presumably rare answer.

Results

Benford's Law

Graph 2 compares the theoretically expected Benford distribution (curves) and the observed distribution (bars) separately for the first and the second digit and the real and the faked interviews (see Graph A1 in the appendix for the digit distributions in the self-administered survey modes).

A first obvious result is that there are considerable deviations from Benford's Law for certain digits in all four subgroups. Looking at the first digit of the real data, it is noticeable that the digits one to three are overrepresented compared to Benford's predictions, while the digits four to nine fall short of the expected percentages. Especially for numbers beginning with the digit two, the observed and expected values diverge strongly, with a difference of almost 15 percentage points. A look at the first digit distribution of the fabricated interviews reveals a similar



Graph 2 Compliance with Benford's Law

picture. The percentage of monetary values that begin with the digit two is even larger than in the real interviews: the observed value is 41% (and thus nine percentage points higher than in the case of the real interviews), while Benford's expectation would range just under 18%. The probability for an initial digit one, on the other hand, is six percentage points below the 30% predicted by Benford.

For the second digit distribution, we find clearly elevated shares of zeros and fives: In the real data, the probability of observing these digits is 14 and 11 percentage points higher than expected, in the fabricated data it is 21 and 14 percentage points above Benford's prediction. In line with the findings documented by some of the previous studies on Benford's second digit (Porras & English 2004; Schräpler & Wagner 2003; Winker et al. 2015), this is clear evidence for rounding. Interestingly, this tendency is slightly more pronounced in the fabricated interviews (although the difference is not statistically significant).

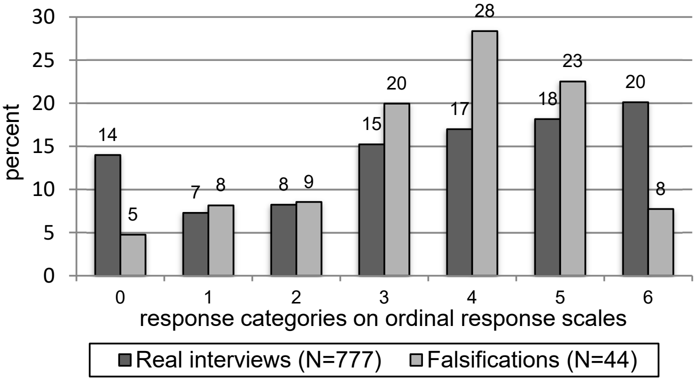
To sum up, neither the authentic nor the fraudulent data followed Benford's Law. Accordingly, chi-squared tests lead to a clear rejection of the hypothesis that the observed data follows Benford's distribution for the four subgroups.

Judging from this finding, we would advise studies with a higher workload per interviewer to consider taking an alternative approach: Instead of comparing falsifications to Benford's Law, each indiviudal interviewers' numerical values could be compared to the empirical sample distribution (as suggested by Swanson, Cho, & Eltinge 2003; Winker 2016). Considering the small number of interviews per interviewer in this data set, we will refrain from looking further into such analyses on the interviewer level.

Extreme Categories on ordinal response scales

As Graph 3 shows, the average number of extreme categories is much lower in the faked data than it is for real interviews. This is true both for the lowest response category 0, which real respondents choose in 14% and falsifiers in 5% of all cases, and for the highest response category 6, which real respondents choose in 20%, and falsifiers in 8% of their answers (both differences in proportions are statistically significant with p<0.01). This means that, while falsifiers underestimate the share of extreme categories by nine to twelve percentage points, the middle categories - and here in particular category 4 - are more frequently chosen in the faked data.[2]

This result is in line with what we find if we sum up the standard deviations within the ordinal response scales for each interview. The average standard deviation for the fake interviews is 1.51, for real interviews 1.94. Comparing the groups in a mean comparison test with cluster-robust standard errors, the difference is statistically significant (p<0.001).



Graph 3 Pooled responses to 17 questions with ordinal response scales

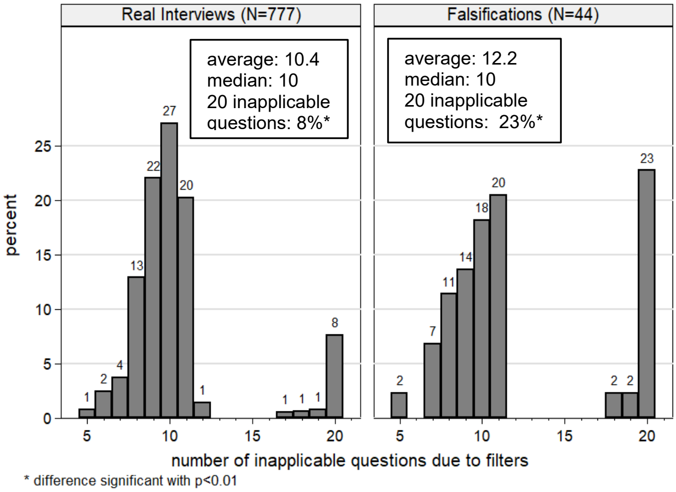
[2] To see if this result was driven by very few falsifiers, the average number of extreme values was individually looked at and compared to the the average number of extreme categories in the real interviews (which was 5.7). Only one out of ten falsifiers (who moreover only faked one interview) ticked a slightly higher number of extreme categories (namely 6). All other falsifiers behaved in line with our hypothesis. Five of them ticked an average of only 2 or fewer extreme categories per interview.

Inapplicable questions in filter branches

We have argued that we would expect falsifiers to use filter branches that shorten the questionnaire. Graph 4 offers evidence that falsifiers indeed pursue this strategy. It shows the absolute number of inapplicable questions, separately for real and fabricated data.

At a first glance, the distributions look similar: The majority of cases have between five and twelve inapplicable questions in filter paths (90% for real interviews and 73% for faked interviews), while a comparatively smaller group avoids 17 or more questions. Differences are particularly evident in this upper part of the distribution. Compared to real interviewers, falsifiers are almost three times as likely to leave all 20 skippable questions unanswered (8% versus 23%; difference in proportions is statistically significant with p<0.01).[3]

The filter path with the largest number of follow-up questions that can be skipped is the one on employment history and working conditions. Interviews can only be in the upper group with 17 to 20 inapplicable questions if the response to the filter question is unanswered or indicates that the respondent has never been in employment. Falsifiers had a strong incentive to answer this question in the negative, because 16 questions on employment were omitted if the respondent had never worked.



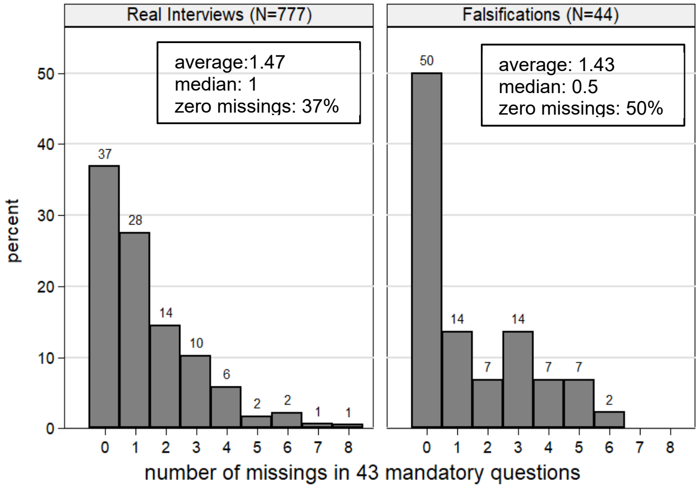
Graph 4 Number of inapplicable questions out of 20 questions in filter branch

[3] As a robustness check, we individually compared each falsifier's average number of inapplicable questions to the average in real interviews to see if results could be driven by very few outliers. 4 out of 10 falsifiers' means ranged between 9.3 and 10, that is slightly lower than the average of 10.4 in real interviews. However, 6 out of 10 falsifiers showed the expected pattern and partly made extensive use of inapplicable questions to presumably shorten the questionnaire.

Item-Nonresponse

Based on the questions that could not be avoided by filters, Graph 5 compares the number of missings in the real and the faked interviews. Our theoretical argument was that falsifiers should avoid missing values in the questions they cannot skip, because too much item-nonresponse might attract the survey agency's attention.

In line with the expectations, falsifiers are more likely to produce questionnaires with zero missing values (50% versus 37%). However, this difference is not reflected in any significant differences, neither in the shares of zero missings nor in the group averages.[4]



Values above 8 (in real interviews) not shown.

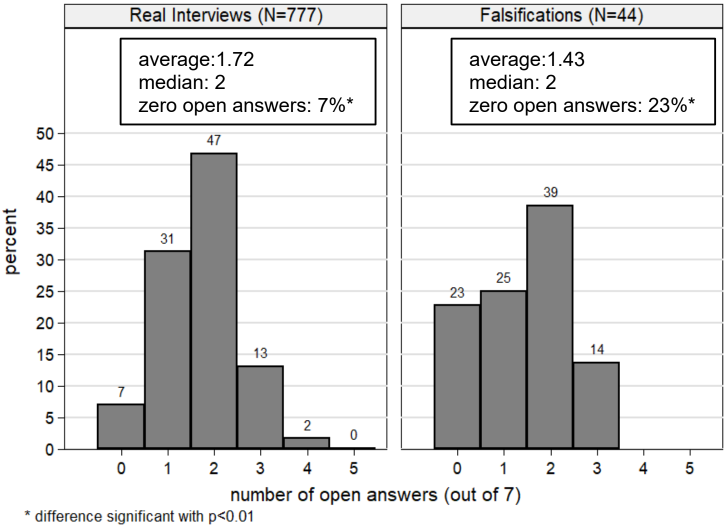
Graph 5 Item-nonresponse in 43 mandatory questions

[4] When looking at the falsifiers' mean numbers of missings, 5 out of 10 yield numbers below the average of real interviews, 4 of them even below 0.2. The other half ranges between 2.2 and 5 missings per interview. These results hint towards heterogeneous interviewer behaviours.

Open answers

Graph 6 shows that there is no significant difference in the mean or median number of open answers between real and fabricated interviews. There is, however, a difference in whether open ended answers are given at all: 7% of real respondents do not give any open answers, compared to 23% of the falsifiers (difference in proportions is significant with p<0.05).[5]

Additional analyses that compared the number of letters in open answers (if there were any) across groups have not revealed any differences in means or medians (see Graph A2 in the appendix for a graphical representation).



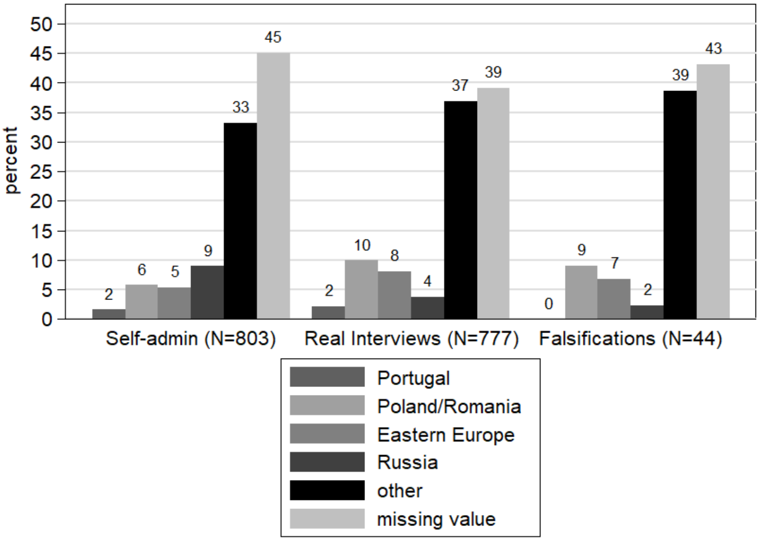
Graph 6 Open answers to seven open questions

[5] 7 out of 10 falsifiers showed lower average numbers of open answers compared to the average in real interviews. Comparing the prevalences of interviews with zero open answers, half of the falsifiers provided at least one open answer per interview. The other half had a likelihood of 29 to 100% to avoid open answers completely.

Trick question

Graph 7 shows the answers to the trick question. In addition to the real and the fabricated personal interviews, the answers given in the self-administered (paper and web) questionnaires are also reported to provide a reference point to a respondent group that could not have been affected by interviewer fraud.[6]

Interestingly and in line with hypothesis 6a, none of the falsified interviews mentioned the correct answer, Portugal. However, it was not a very frequent answer in the real interviews or in the self-administered questionnaires. The prevalence of Poland, Romania or other Eastern European countries did not differ enough (or in the expected direction) across groups to serve as an indicator for fraud in the face-to-face sample. We therefore reject hypothesis 6b.



Graph 7 Answers to trick question on income inequality in Europe across modes

[6] Recruitment into the two survey modes took place by drawing two separate random samples of the residential population of Germany (more information in Sauer et al. 2010). For the purposes at hand, we are implicitly assuming that the random differences between samples do not affect the answers to the trick question.
As one of the reviewers pointed out, self-administered questionnaires might still contain falsifications if the survey agency added invented cases. Although this is a theoretical possibility, we have no reason whatsoever to believe that this happened in our data.

As expected, the answer Portugal predicts perfectly that an interview is authentic. Since the other answers do not help to distinguish between real and faked interviews, the trick question will not be further analyzed in the following multivariate analyses. It will, however, be included as one of the open questions, which can either be answered or missing.

Multivariate analyses

To examine the relative importance of the different indicators, the next step is to use them as potential predictors for fraud in a logistic regression model. The dichotomous dependent variable indicates whether an interview is real (coded as 0) or a case of known fraud (coded as 1).

Table 1 summarizes all explanatory variables. For the indicators on *extreme response categories* and *item-nonresponse*, some categories at the upper end of the scale were grouped together due to small numbers of cases.

Table 1 Explanatory variables in regression model

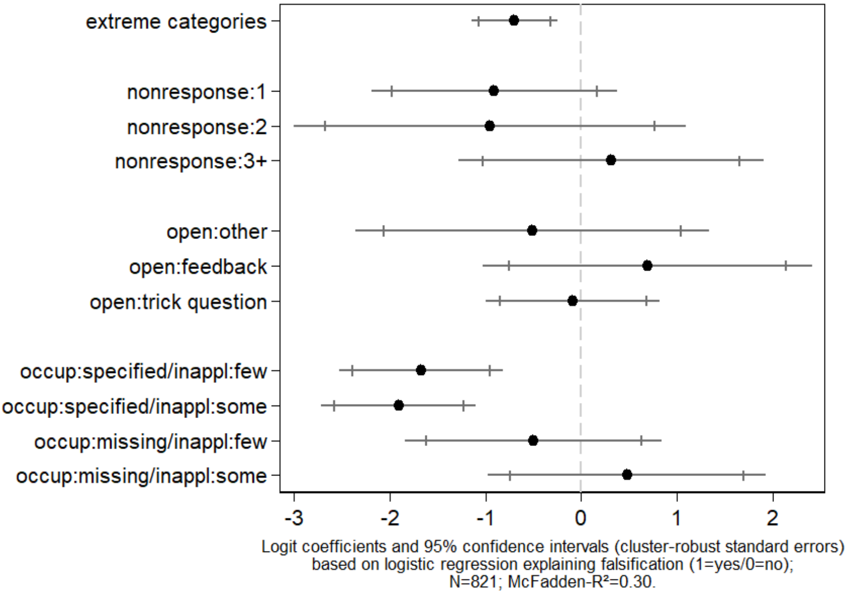
Variable	Concept and Coding	
extreme categories	number of extreme categories on 17 items with ordinal response scales (continuous variable)	→ hypothesis 2
item-nonresponse	number of missings in mandatory/unfiltered questions (0/1/2/3+, specified as dummies)	→ hypothesis 4
open: other	answered any “other, please specify” category (0/1)	→ hypothesis 5
open: feedback	answer to feedback question at the end of the survey (0/1)	→ hypothesis 5
open: trick question	answer to question on income inequality in Europe (0/1)	→ hypothesis 5
occupation * inapplicable	open question on occupation (skipped / specified / missing) number of inapplicable questions (5-9 “few”/ 10-12 “some”/ 16-20”many”) 0: occupation: skipped / inapplicable: many 1: occupation: specified / inapplicable:few 2: occupation: specified / inapplicable: some 3: occupation: missing / inapplicable: few 4: occupation: missing / inapplicable: some	→ hypothesis 5 → hypothesis 3

There is a peculiarity in the data concerning the open question that elicits the respondent's current (or alternatively last) *occupation*. It is positioned within the longest filter branch of the questionnaire that was designed to collect information on the respondent's employment and working conditions. Failure to specify an occupation can thus stem from item-nonresponse or from skipping the entire filter branch. A high number of *inapplicable questions* essentially indicates that the complete filter branch on employment, including the item on occupation, was skipped. In terms of coding, this correlational pattern is reflected in the interaction of two categorical variables, the respondent's reaction to the occupation question (skipped / missing / specified) and the number of inapplicable questions (few / some / many). The reference category indicates an inapplicable occupation question and a generally high number of inapplicable questions due to filters throughout the questionnaire.

Apart from respondent occupation, all other *open answers* are captured by dichotomous variables. Since very few respondents specified additional information in "other, please specify" categories, these questions were subsumed into one dummy variable indicating if any (versus no) open answer was provided.

Graph 8 shows the results of the logistic regression model (see Table A1 in the appendix for the full regression table). A first result is that there is a significant negative correlation between the number of extreme categories in the questionnaire and the log-odds that an interview was fabricated ($p<0.01$). This means that hypothesis 2 is clearly confirmed and falsifiers indeed avoided extreme answers on ordinal response scales. Equally clearly, we cannot find any evidence that item-nonresponse is related to fraud (hypothesis 4). Regarding hypothesis 5, it is interesting to note that the absence of open questions (or the numbers of letters in open answers as mentioned before) does not generally come with higher probabilities of falsification. However, if we look at the question on respondent occupation, questionnaires that contained a job description were significantly less likely to be falsifications than interviews with item-nonresponse ($p=0.08$ for *few* inapplicable questions, $p>0.01$ for *some* inapplicable questions) or a skipped question ($p<0.01$ for *few* and *some* inapplicable questions).

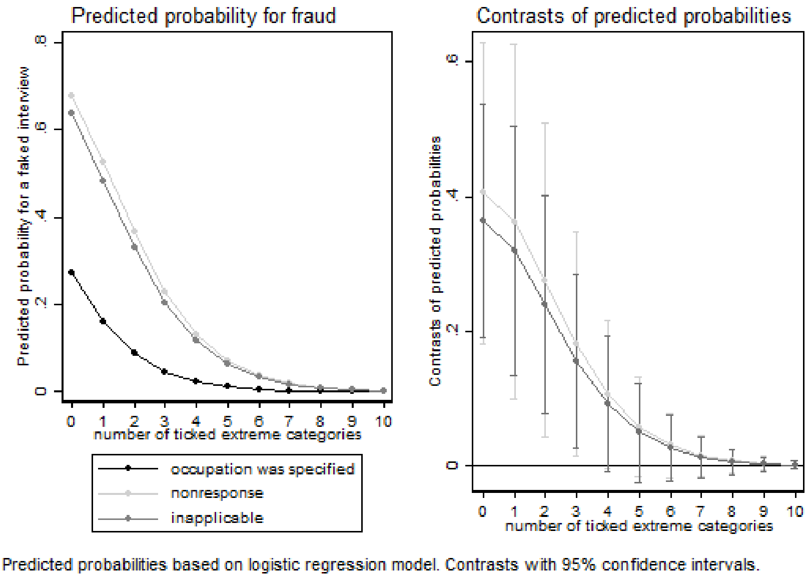
When it comes to inapplicable questions (hypothesis 3), regression coefficients for respondents that got filtered over *few* or *some* questions were very similar. Only a high number of skipped questions (including the occupation question) was predictive for fraud in our data. This pattern somewhat suggests that the intention to skip the open question on occupation was the dominant one compared to finding the shortest path through the filter paths of the questionnaire.



Graph 8 Potential Indicators for Fraud (logistic regression)

Looking into descriptives, falsifiers report around three times as many respondents who have never been in employment and thus automatically skip the 16 additional questions on employment (27.3 versus 9.5%). Among those who are asked about their current or previous occupation, 93.0% of the real interviews provide a job title, while only 81.3% of the falsifications do. With these considerable differences in response patterns, the question on the respondent's occupation was very helpful in identifying fabricated interviews. Part of the explanation surely lies in how easy it would be to check the correctness of the answer. Respondents certainly remember what they do or did for work, and even other household members could presumably provide the right answer, e.g. in a follow-up check by phone. However, the question is not only easily verifiable but also easily avoidable by a strategic use of the filter paths. The combination of both peculiarities seems to explain its success in identifying fraud.

Graph 9 translates the two significant indicators from the regression model into a visual illustration: It shows how the number of extreme categories and the



Graph 9 Predicted probability for fraud dependent on the strongest predictors

question on occupation interact.[7] The y-axis of the graph on the left side shows the predicted probability that an interview has been fabricated.

At a first glance, the graph shows that the effect of one variable heavily depends on the other variable. Generally speaking, the probability of fraud decreases with an increasing number of selected extreme categories on the ordinal response scales of the questionnaire. This is even more so if the open question on the respondent's occupation has not been answered - due to nonresponse or because the respective filter branch was skipped entirely.

Within the groups of interviews that did not indicate any occupation, the predicted probability of fraud can rise to a maximum of 63% or 68% if zero extreme categories have been ticked. For interviews in which the occupation was specified, the slope is less steep and only climbs to 27% if we move towards the lower end of the x-axis indicating low numbers of extreme categories. Comparing the interviews with and without the open answer, the predicted probability of fraud hardly

[7] Graph 9 only refers to occupation rather than the interaction between occupation and inapplicable questions. This is justified by the fact that there weren't any significant differences between the groups with few and some inapplicable questions in the previous model. In addition, the number of cases with a missing value for the occupation question seemed too small (N=55) to identify further subgroups (as can be seen by the large confidence intervals for the last two regression coefficients in Graph 8).

changes at the upper end of the scale, while the open answer accounts for a change of 36 to 41 percentage points if no extreme category was chosen.

The illustration on the right side of Graph 9 treats interviews with a specified occupation as the reference category (straight line) and compares it to the two groups without valid answers (curves). The confidence intervals indicate that the group differences between the interviews with and without a specified occupation are significant if zero to three extreme categories were ticked and insignificant if four or more extreme categories were ticked.

Judging by the maximum changes in predicted probabilities that can be attributed to the variables, the results suggest that the number of extreme categories is an even stronger indicator of fraud than failing to provide a job title in a filter branch that falsifiers might want to skip.

Related findings: standard deviations, acquiescence and social desirability

Theoretically, we could have considered standard deviations in the ordinal response scales or the amount of straight lining (as done by Blasius & Thiessen 2018) instead of the number of extreme categories. Empirically, however, the number of ticked extreme categories was more strongly correlated with fraud and the standard deviations did not add any explanatory power to the regression models once the model controlled for extreme categories.

Following Kemper & Menold (2014), we tested two more alternatives to examining the extreme categories, namely acquiescence (in both ordinal response scales) and socially desirable answers (in the social desirability scale). In their paper on laboratory falsifications, Kemper & Menold (2014) report that falsifiers provide “overly positive self-descriptions” (p. 96) when asked about socially desirable behaviors as well as a higher tendency to acquiesce irrespective of question content. These results were not replicable with our data. Both options performed worse as indicators for fraud than the number of extreme categories. Generally, adding indicators for self- and other-deception did not help the predictive power of the regression model. Descriptively, falsifiers tended to give negative self-descriptions in five out of six cases. Curiously, they were less likely to agree to the statement “I am always honest to others”. Although the difference between groups was not statistically significant, this finding would be in line with the argument that falsifiers use themselves as a reference point when fabricating data (see Landrock 2017).

Predicting fraud based on internal indicators alone?

In response to a reviewer comment, Table 2 presents the numbers of falsifications that would have been correctly classified. In other words, we are treating our data as a case of supervised learning relying on internal indicators (on interview level)

Table 2 Confusion Matrix

	Actual: Real Interview	Actual: Falsification	Total
Prediction: real	757	22	779
Prediction: falsified	20	22	42
Total	777	44	821

alone. As the analyses are carried out on the interview level, we are allowing for the situation that interviewers only faked some of their assigned interviews.

To predict fraud based on the logistic regression model, it makes sense to work with an educated guess about the expected share of falsifications instead of using the 50% mark as a cut-off point. Similar to ordinary clustering approaches, the method would otherwise be very likely to identify an unrealistically high number of falsifications (De Haas & Winter 2016). In this case, the upper 5% of interviews that were most suspicious from internal indicators are treated as falsifications.

The results suggest that only by taking internal indicators at interview level into account, we could have identified 22 out of 42 confirmed falsifications correctly. 20 would have remained undetected and there would have been 22 new suspicious cases (which could have been subjected to further checks if the internal indicators had been part of the quality control during fieldwork).

This little exercise highlights a point that has been made before: Internal indicators seem to work well to identify some falsifications but not others (a finding that is confirmed by Thissen & Myers 2016). As a consequence, it is not reasonable to rely only on one type of indicator. Instead it should be the goal to combine as many as possible. In settings where the number of conducted interviews per interviewer is higher, internal indicators on interviewer level can supplement the internal indicators on interview level that we used in this study. These could include duplicate checks (Koczela, Furlong, McCarthy, & Mushtag 2015), analysis of similar response patterns in ordinal response scales (Blasius & Thiessen 2015), and comparisons of so-called "content-related patterns", contrasting interviewer-specific and overall sample means (Kosyakova, Olbrich, Sakshaug, & Schwanhäuser 2019; Weinauer 2019).

To some extent (although in varying degrees), all internal indicators have the disadvantage that falsifiers will be able to adjust their response style if they are aware that a certain kind of check is in place (Winker 2016). However, the idea is not that one method will identify all falsifications, but that more checks will heigthen the chance of being detected and will make fraud less attractive to interviewers. As Thissen & Myers (2016) put it: “Each method can be circumvented, but a combination of methods acts as a series of barriers, and patterns of falsification that might slip past one type of review may be caught by another.”

Conclusion

This study provides empirical insights into the response patterns of falsifiers. It contributes to a wider body of literature aiming to develop more efficient monitoring strategies to prevent interviewer fraud and potential bias in surveys. For this purpose, we examined the potential of statistical ex-post analysis of response patterns for the identification of fabricated data. In contrast to the wide majority of studies on interviewer fraud, we did not draw on laboratory falsifications but could rely on authentic cases of interviewer fraud stemming from a survey project on income inequality in Germany. Out of 821 face-to-face interviews, 44 were identified as falsifications by external control mechanisms and admitted as such by the survey agency. These data were particularly suitable to test the potential of internal indicators for fraud detection: To what extent do the attempts of falsifiers to produce unsuspicious data lead to recognizable response patterns?

Drawing on somewhat contradictory empirical evidence from previous studies, hypotheses were formulated as to how rationally acting falsifiers would navigate through a questionnaire. Real and faked interviews were then compared with regard to various testable criteria, namely the number of selected extreme categories in ordinal response scales, answers to open-ended questions, item-nonresponse, strategic use of filter questions to shorten the questionnaire, and the compliance of reported numbers with Benford‘s Law. In addition, we reported results from an innovative trick question.

In the multivariate analysis, the avoidance of extreme categories on ordinal response scales proved to be the strongest indicator for fraud. This approach also outperformed alternative indicators, namely socially desirable answers and acquiescence in the ordinal response scale. Apart from that, missing data in one particular open-ended question significantly predicted fraud: the open question on the respondent’s current or previous occupation. Missing values could occur either because the respondents refused to provide a job title or because they had never been in employment, in which case the interviewer was supposed to skip a filter branch of 16 questions on employment. None of the other open-ended questions, however, helped to detect fraud, and neither did the number of letters in open answers. This means that fewer or shorter open answers per se do not seem to be suitable indicators. Similarly, questionnaires with few and moderate numbers of inapplicable questions did not differ in their probability of being falsified. Only a high number of inapplicable questions (stemming from skipping the entire filter branch on employment) was strongly correlated with fraudulent data. Although we cannot fully disentangle the effect of not answering an easily verifiable open-ended question and using filters to shorten the questionnaire, we can definitely say that the combination of both successfully identified fraud in our case.

Another interesting finding of the above analyses is that, contrary to the original hypothesis, falsifiers and real respondents did not differ in their shares of item-nonresponse. Apart from that, neither the implemented trick question nor Benford's Law were helpful in detecting fraud. Although all falsifiers avoided the rare correct answer to the trick question, very few real respondents gave the correct answer. When comparing the first and the second digits of reported monetary values to Benford's Law, this criterion proved to be highly problematic, since even the basic assumption that the authentic interviews should approximately follow the Benford distribution was violated. Judging from this finding, it seems more promising to experiment with rounding behavior (although differences failed to reach a significant level in our data) or to compare individual clusters of interviews to the average empirical distribution (as suggested by Swanson, Cho, & Eltinge 2003; Winker 2016). This approach can be pursued if the number of interviews per interviewer is sufficiently high to allow analyses at the interviewer level.

A limitation of this study is that we cannot rule out that, despite extensive checks, further unnoticed falsifications have remained in the data. Comparing the answers to the trick question and the digits from the monetary values to the questionnaires from the self-administered survey modes shows that this might be the case, although differences between real face-to-face interviews and self-administered questionnaires could in principal also stem from imperfect randomisation, mode differences in non-response bias or response behavior. Despite these uncertainties, the present study could show that more than half of our authentic cases of fraud which were detected by external control mechanisms would also have raised suspicion in a statistical ex-post analysis of their response patterns. This result can encourage researchers and survey agencies to use both types of indicators in combination when identifying suspicious cases.

One possible practical approach could be to run checks on response patterns more continuously during the fieldwork period. In line with AAPOR recommendations (AAPOR 2003), it is desirable to complement random picking of interviews with a more targeted strategy for additional quality control. This is where statistical approaches could help to identify suspicious cases for further checks. Ideally such a routine would already be at work during fieldwork, not after its completion. A corresponding statistical routine would have to identify outliers based on a set of internal indicators. This could be done by assigning suspicion points to interviews, by generating pareto charts of the collected paradata on a weekly basis (as suggested by Gwartney 2013). Concrete advise for implementing statistical routines that identifiy suspicious interviewers who deviate significantly from the overall sample mean has been provided by Weinauer (2019) and Kosyakova et al. (2019) who draw on the concept of "meta-indicators".

Regarding the selection of internal indicators for statistical procedures, the presented results suggest that it is worth looking at the peculiarities of the specific questionnaire to e.g. identify high-risk questions that are easily verifiable or crucial for long filter paths. A general recommendation is to make use of as many indicators as possible - internal and external ones - to identify and substantiate suspicions.

References

AAPOR (2003). Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects. American Association for Public Opinion Research (AAPOR). https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf

Bauer, J., & Gross, J. (2011). Difficulties Detecting Fraud? The Use of Benford's Law on Regression Tables. *Jahrbücher für Nationalökonomie und Statistik*, 231(5+6).

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society,* 78(4), 551–572.

Blasius, J., & Thiessen, V. (2018). Perceived Corruption, Trust, and Interviewer Behavior in 26 European Countries. *Sociological Methods & Research* (online first).

Blasius, J., & Thiessen, V. (2015). Should We Trust Survey Data? Assessing Response Simplification and Data Fabrication. *Social Science Research* 52, 479–493.

Blasius, J., & Thiessen, V. (2013). Detecting Poorly Conducted Interviews. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 67-88). Peter Lang Academic Research.

Bredl, S., Winker, P., & K**ö**tschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology*, 38(1), 1–10.

Brüderl, J., Huyer-May, B., & Schmiedeberg, C. (2013). Interviewer Behavior and the Quality of Social Network Data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 147–160). Peter Lang Academic Research.

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*, 316–320.

Crespi, L. (1945). The Cheater Problem in Polling. *Public Opinion Quarterly* 9, 431-445.

De Haas, S., & Winker, P. (2016). Detecting Fraudulent Interviewers by Improved Clustering Methods – The Case of Falsifications of Answers to Parts of a Questionnaire. *Journal of* Official Statistics, 32(3), 643–660.

De Haas, S., & Winker, P. (2014). Identification of Partial Falsifications in Survey Data. *Statistical Journal of the IAOS* 30(3), 271–281.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics*, 34(3), 321–329.

Diekmann, A., & Jann, B. (2010): Benford's Law and Fraud Detection. Facts and Legends. *German Economic Review* 11(3), 397–401.

Dorroch, H. (1994). *Meinungsmacher-Report. Wie Umfrageergebnisse entstehen.* Göttingen: Steidl.

Esser, H. (1999). *Die Wert-Erwartungstheorie. Soziologie - Spezielle Grundlagen, Band I: Situationslogik und Handeln.* Frankfurt am Main: Campus Verlag, Kapitel 7.

Finn, A., & Ranchhod, V. (2017). Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey. *The World Bank Economic Review,* 31(1), 129–157.

Gwartney, Patricia A. (2013). Mischief versus Mistakes: Motivating Interviewers to Not Deviate. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 195–215). Peter Lang Academic Research.

Hood, C., & Bushery, J. (1997). Getting more bang from the reinterviewer buck: Identifying "at risk" interviewers. *Proceedings of the American Statistical Association*, 820–824.

Josten, M., & Trappmann, M. (2016). Interviewer Effects on a Network-Size Filter Question. *Journal of Official Statistics,* 32(2), 349–373.

Kemper, C. J., & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* 10(3), 92–99.

Koch, A. (1995). Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994. *ZUMA Nachrichten,* 36, 89–105. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-208984

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). Curbstoning and beyond: Confronting Data Fabrication in Survey Research. *Statistical Journal of the IAOS* 31(3), 413–422.

Kosyakova, Y., Olbrich, L., Sakshaug, J., & Schwanhäuser, S. (2019). Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany. FDZ-Methodenreport 02.2019. http://doku.iab.de/fdz/reporte/2019/MR_02-19_EN.pdf

Kroneberg, C., & Kalter, F. (2012). Rational Choice Theory and Empirical Research: Methodological and Theoretical Contributions in Europe. *Annual Review of Sociology,* 38(1): 73–92.

Landrock, U. (2017). How Interviewer Effects Differ in Real and Falsified Survey Data: Using Multilevel Analysis to Identify Interviewer Falsifications. *methods, data, analyses,* 11(2), 163-188.

Menold, N., & Kemper, C. J. (2014). How Do Real and Falsified Data Differ? Psychology of Survey Response as a Source of Falsification Indicators in Face-to-Face Surveys. *International Journal of Public Opinion Research,* 26(1), 41–65.

Menold, N., Landrock, U., Winker, P., Pellner, N., and Kemper, C. J. (2018). The Impact of Payment and Respondents' Participation on Interviewers' Accuracy in Face-to-Face Surveys: Investigations from a Field Experiment. *Field Methods*, 30(4), 295–311.

Menold, N., Winker, P., Storfinger, N., & Kemper, C. (2013). A Method for Ex-Post Identification of Falsifications in Survey Data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 25–47). Peter Lang Academic Research.

Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O., & Hsieh, Y. P. (2016). Interviewer falsification: Current and best practices for prevention, detection, and mitigation. *Statistical Journal of the IAOS* 32, 313–326.

Nelson, J. E., & Kiecker, P. L. (1996). Marketing Research Interviewers and Their Perceived Necessity of Moral Compromise. *Journal of Business Ethics* 15, 1107–1117.

Nigrini, M. (1999). I've got your number. *Journal of Accountancy* 187(5), 79–83.

Porras, J., & English, N. (2004). Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys. *Proceedings of the American Statistical Association*, 4223–4228.

Reuband, K.-H. (1990). Interviews, die keine sind. „Erfolge" und „Misserfolge" beim Fälschen von Interviews. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie: KZfSS,* 42, 706–733.

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2010). Konzeption und Durchführung der Studie „Einkommensgerechtigkeit in Deutschland" im Rahmen des Projekts „Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen". Feldbericht.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., & Wagner, G. G. (2005). Automatic Identification of Faked and Fraudulent Interviews in the German SOEP. *Schmollers Jahrbuch: Journal of Applied Social Science Studies,* 125(1), 183–193.

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

Schr**ä**pler, J.-P., & Wagner, G. G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys. An analysis by means of genuine fakes in the raw data of SOEP. IZA Discussion Paper No. 969. http://ftp.iza.org/dp969.pdf

Schräpler, J-P. (2011). Benford's Law as an Instrument for Fraud Detection in Surveys Using the Data of the Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik* 231.

Schreiner, I., Pennie, K., & Newbrough, J. (1988). Interviewer Falsification in Census Bureau Surveys. *Proceedings of the American Statistical Association*, 491–496.

Schwanhäuser, S., Sakshaug, J., Kosyakova, Y., & Kreuter, F. (2020). Statistical Identifcation of Fraudulent Interviews in Surveys: Improving Interviewer Controls. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter & B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (pp. 91-105). Chapman and Hall/CRC.

Storfinger, N., & Winker, P. (2013). Assessing the Performance of Clustering Methods. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 49-65). Peter Lang Academic Research.

Swanson, D. Cho, M, J., & Eltinge, J. (2003). Detecting Possibly Fraudulent or Error-Prone Survey Data Using Benford's Law. *Proceedings of the American Statistical Association*, 4172–4177.

Thissen, R. (2014). Computer Audio-Recorded Interviewing as a Tool for Survey Research. *Social Science Computer Review* 32(1), 90–104.

Thissen, M. R., & Myers, S. K. (2016). Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality. *Statistical Journal of the IAOS* 32(3), 339–47.

Wagner, J., Olson, K., & Edgar, M. (2017). The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata. *Survey Research Methods* 11, 218-233.

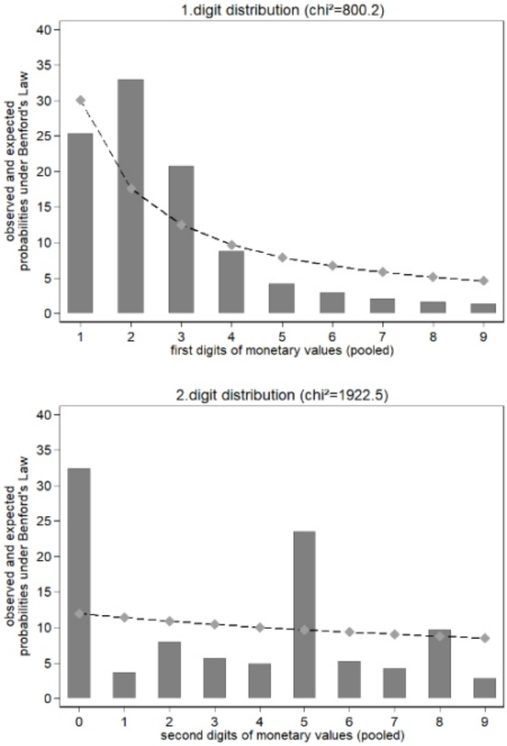
Weinauer, M. (2019). Be a Detective for a Day: How to Detect Falsified Interviews with Statistics. *Statistical Journal of the IAOS* 35(4), 569–75.

Winker, P. (2016). Assuring the Quality of Survey Data: Incentives, Detection and Documentation of Deviant Behavior. *Statistical Journal of the IAOS 32*(3), 295–303.

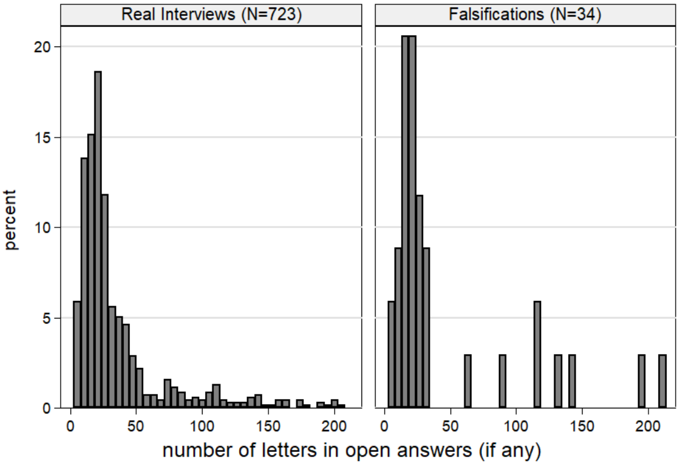
Winker, P., Kruse, K.-W., Menold, N., & Landrock, U. (2015). Interviewer Effects in Real and Falsified Interviews: Results from a Large Scale Experiment. Statistical Journal of the IAOS 31(3), 423–434.

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). Journal of Individual Differences 34(1), 32–40.

Appendix A



Graph A1 Compliance with Benford's Law in the self-administered surveys (N=803, 2612 digits)



Values above 215 digits (in real interviews) not shown.

Graph A2 Number of letters in open answers (if any)

Table A1 Potential indicators for fraud – regression table

Logistic Regression explaining falsification (1=yes/0=no) (logit coefficients; cluster-robust standard errors)	
extreme categories	-0.70**
	(-3.07)
nonresponse:1	-0.91
	(-1.39)
nonresponse:2	-0.96
	(-0.92)
nonresponse:3+	0.31
	(0.38)
open:other	-0.51
	(-0.54)
open:feedback	0.69
	(0.79)
open:trick question	-0.09
	(-0.19)
occup:specified/inappl:few	-1.67***
	(-3.83)
occup:specified/inappl:some	-1.91***
	(-4.63)
occup:missing/inappl:few	-0.50
	(-0.73)
occup:missing/inappl:some	0.48
	(0.64)
intercept	1.05
	(1.67)
McFadden Pseudo-R^2	0.31
N	821

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Appendix B. Question Wording by Indicators

(for the original questionnaire in German language, see Sauer et al. 2010)

- **Nonconformity with Benford's Law**

What do you think, what is the average gross income per month for a full time position in Germany?
(This is the amount that employees receive from their employer before taxes and social security contributions are deducted.)
About __________ ,- Euro per month

What is your own monthly gross income from your employment?
(This is the amount that you receive from your employer before taxes and social security contributions are deducted. If you are self-employed, please fill in the average of what you earn per month.)
About __________ ,- Euro per month

In case you don't perceive your own income as fair, what would be a fair monthly gross income for you?
About __________ ,- Euro per month

What is the monthly net income of your household overall?
(Please sum up all types of household income, including your own. This also includes income from rentals and royalties, pensions, unemployment, social security and child benefits, rent subsidy and other types of income.)
About __________ ,- Euro per month

To cover for your running expenses, what is the minimum monthly net income that your household would need?
(Please specify the amount you need per month to cover for housing, food, clothes, heating and your personal basic needs.)
About __________ ,- Euro per month

- **Avoiding extreme categories**

In your opinion, to what extent should the following factors matter for a fair gross income?

Age ……………………………….. not at all 0-1-2-3-4-5-6 very much

Sex ……………………………….. not at all 0-1-2-3-4-5-6 very much

Education ………………………… not at all 0-1-2-3-4-5-6 very much

Number of children ………………. not at all 0-1-2-3-4-5-6 very much

Job ………………………………... not at all 0-1-2-3-4-5-6 very much

Job experience …………………… not at all 0-1-2-3-4-5-6 very much

Health condition ………………….. not at all 0-1-2-3-4-5-6 very much

Time working for company ……… not at all 0-1-2-3-4-5-6 very much

Size of company …………………. not at all 0-1-2-3-4-5-6 very much

Economic situation of the company not at all 0-1-2-3-4-5-6 very much

Performance on the job …………... not at all 0-1-2-3-4-5-6 very much

To what extent do you agree with the following statements?

The first impression I have of people usually
turns out to be right...…………………………… not at all 0-1-2-3-4-5-6 very much

I am usually very sure of my judgements...…… not at all 0-1-2-3-4-5-6 very much

I am not always aware of the reasons for my
actions...……………………………………….. not at all 0-1-2-3-4-5-6 very much

It has happened that I kept too much change..... not at all 0-1-2-3-4-5-6 very much

I am always honest with other people…………. not at all 0-1-2-3-4-5-6 very much

I have never taken advantage of somebody….... not at all 0-1-2-3-4-5-6 very much

- **Avoiding open answers**

What is your current job, or what was your last job? Please specify the job title and describe your role exactly.

What do you think, which European country does currently have the highest income inequality?

We might have missed something that you consider important. Is there anything you want to add or comment on?