International Journal of Public Opinion Research Vol. 26 No. 1 2014

doi:10.1093/ijpor/edt017 Advance Access publication 27 June 2013

How Do Real and Falsified Data Differ? Psychology of Survey Response as a Source of Falsification Indicators in Face-to-Face Surveys

Natalja Menold[1] and Christoph J. Kemper[2]

[1]GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany and [2]Institute for Medical and Pharmaceutical Proficiency Assessment, Mainz, Germany

Abstract

Interview falsifications in face-to-face surveys have the potential to seriously contaminate survey data. Thus, detection methods are of crucial importance. One idea is to use several falsification indicators derived from existing survey data. Here, real and falsified data are compared in two experimental studies. Response sets are used to develop formal indicators, and content of responses is used to develop content-related indicators. During the analysis, numerous formal indicators are tested and compared with content-related indicators. The results indicate that the majority of formal indicators allow for differentiation between real and false data, and, in addition, certain content-related indicators are useful for such differentiation. The results are discussed in terms of the usefulness of indicators in identifying potential falsifiers.

Introduction

Despite a growing volume of online surveys and the wide use of telephone surveys, the face-to-face interview remains an important mode of data collection, and it is still considered to be a preferable survey mode owing to the associated advantages concerning coverage and data quality (e.g., Groves et al., 2004). The interviewer plays a central role in face-to-face surveys with a non-negligible positive impact, for example in obtaining participation or in providing help so that survey questions are correctly understood. The interviewer

All correspondence concerning this article should be addressed to Natalja Menold, PO Box 12 21 55, D-68072 Mannheim, Germany. E-Mail: natalja.menold@gesis.org.

can also have a potentially negative influence, for example in cases of mistakes, low accuracy, or even falsifications. AAPOR (2003, p. 1) defines interviewer falsification as "the intentional departure from the designed interviewer guidelines or instructions, which could result in the contamination of data." Most severe are falsifications in which interviewers themselves produce the data instead of obtaining it from the sampled person. In our research, we focus on such interview falsifications to test indicators in terms of usability for detecting false data or putative dishonest interviewers (interviewers "at risk"; Hood & Bushery, 1997).

There are two kinds of interview falsifications—falsification of entire interviews and falsification of a part of the interview (partial falsifications). Partial falsifications are assumed to be more common than entire falsifications (Evans, 1961). In the case of partial falsifications, interviewers obtain basic data from sampled persons during a so-called short interview. Having this basic data then makes the subsequent fabrication of remaining data easier. Basic data includes, for example, size of the household or other demographic information. Only a small amount of evidence exists in the literature about the prevalence of interview falsifications. According to a few authors, falsifications may make up between <1% and up to 5% of interviews (Biemer & Stokes, 1989; Case, 1971; Koch, 1995). In western countries, using well-established control procedures, the overall prevalence of falsified interviews in face-to-face surveys is reported to be low (cf. AAPOR, 2003). However, fabricated interviews are not always reliably detected. The commonly used control procedure of sending postcards is rather ineffective particularly in detecting partially falsified interviews, owing to the often rather low willingness of participants to respond, memory problems, and the limited amount of information that can be confirmed. Thus, it is reasonable to doubt that the prevalence reported in the literature may be underestimated, especially for surveys regarding sensitive topics (Turner, Gribble, Al-Tayyib, & Chromy, 2002) or surveys conducted in developing countries where rigorous control procedures are not always applied (Bredl, Kötschau, & Winker, 2012). It was also shown that even a small number of falsifications may seriously affect the results of multivariate analyses (Schraepler & Wagner, 2003). The authors showed that the inclusion of fabricated German Socio Economic Panel data in a multivariate regression reduced the estimated effect on log gross wages by ~80%, even though the proportion of fabricated interviews was <2.5%. Consequently, methods that help to detect data falsifications are of crucial importance for survey research. The most reliable tool for identification of falsifications is the re-interview (Bredl, Storfinger, & Menold, 2013). However, re-interviews, like postcards, are susceptible to memory effects and are simply not practical owing to high costs and effort. Thus, methods that allow for easier evaluation of interviewers' work are needed. One idea is to use data analysis methods applying several falsification indicators, which help to identify interviewers "at risk"

and to conduct more focused control procedures instead of randomly selecting interviewers for controls (Bredl, Storfinger, & Menold, 2013). Different data sources can be used to define falsification indicators for detection methods. We differentiate between three kinds of data sources:

1. Meta-data and para-data are data related to the process of data collection, for example implausible success rates (Turner, Gribble, Al-Tayyib, & Chromy, 2002), or time stamps in computer-assisted interviews (Bushery, 1999; Murphy, Baxter, Eyerman, Cunningham, & Kennet, 2004).
2. The content of responses to questions is the source of "content-related" indicators. Murphy, Baxter, Eyerman, Cunningham, and Kennet (2004) used responses related to lifetime consumption of different drugs in the American National Drug Survey on Drug Use and Health. With an ex-post analysis the authors found higher reporting of drug consumption in falsified data than in non-falsified data (here falsifiers were known a priori). In another survey on a sensitive topic (sexually transmitted diseases), falsifiers, who were also known a priori, reported that their "survey participants" were more sexually active than real respondents (Turner, Gribble, Al-Tayyib, & Chromy, 2002). Next, Benford's law (Benford, 1938) was used to distinguish between real and false data (e.g., Schraepler & Wagner, 2003; Wang & Pedlow, 2005; Porras & English, 2004). When applying Benford's law, it is assumed that real metric data are distributed with an unequal probability of a number (from 1 to 9) being a first digit. However, it was shown that real survey data deviate from Benford's distribution, but distribution of first or second digits in falsified data sometimes differs from real distributions, leading some authors to use these differences as indicators of falsification (Porras & English, 2004). Hence, Porras and English discussed limiting this method, as real respondents often use round numbers (multipliers of 5 or 10) when providing numerical answers (rounding, see later in the text).

   Studies conducted in Germany (Hippler, 1979; Reuband, 1990; Schnell, 1991) systematically compared results from real and false data by using the following experimental design. First, a subsample was drawn from an existing survey (e.g., German General Social Survey, ALLBUS). Then, false data were produced for this subsample by students or researchers playing the role of falsifiers. Before this falsifying process, some basic data about real survey participants had been provided to the falsifiers. Data produced by the falsifiers were then compared with real survey data. The studies showed that falsifiers were able to produce opinion and behavior-related data that were close to real distributions. Hence, falsifiers' precision in predicting responses of real respondents may depend on the domain of falsification, for example whether it is a

matter of behavior or opinions, as well as what basic information is known to falsifiers (Hippler, 1979).

3. Differences in response behavior form the third source of indicators. This does not mean differences in the content of responses, but rather in item response and non-response, as well as the way in which a response is given, for example different response sets. We call the corresponding indicators "formal" indicators. Schraepler and Wagner (2005) used extreme responding, variance in responding and item non-response (INR) to identify (previously detected) falsifiers in the Socio Economic Panel. Bredl, Kötschau, and Winker (2012) used responses to open-ended questions, semi-open-ended questions, filter questions, and also the frequency of extreme responses as indicators for their ex-post analysis of real survey data to identify dishonest interviewers. Information from these different indicators was exploited by means of cluster analyses to classify interviewers, which resulted in high selectivity in terms of identifying interviewers "at risk." Even though this method appears to show good results, it also delivered a considerable number of false positives. However, this method can be improved if more powerful indicators are available.

In our research, we focus on content-related and formal indicators, which can be used in combination with meta-data or para-data to more effectively select interviewers for re-interview during control procedures. In addition, these indicators can be used as part of detection methods to evaluate existing data. Formal indicators might have advantages over content-related indicators. First, in the case of content-related indicators, differences in data distributions between real and false data may be marginal (Reuband, 1990; Schnell, 1991); second, these might depend on the basic data known to falsifiers (Hippler, 1979), or third, questions on sensitive topics (drug consumption, sexual behavior) should be included in the questionnaire (Murphy, Baxter, Eyerman, Cunningham, & Kennet, 2004; Turner, Gribble, Al-Tayyib, & Chromy, 2002). In contrast, differences with respect to formal indicators might be larger and independent of the topic of questions or information known to potential falsifiers. However, we do not know of studies that compare formal and content-related indicators. Thus, when using formal indicators, one cannot be sure about using the best ones, in other words, those that are most sensitive to falsifications.

Our first research purpose is to test formal indicators used by other authors and to find additional indicators that are sensitive to falsifications. Our second research purpose is to compare formal indicators with content-related indicators to identify the most effective ones for revealing differences between real and false data. In the next section, we provide a further classification of

formal and content-related indicators, which we use for a specification of our research questions.

A Classification of Falsification Indicators

Formal Indicators

To classify indicators grounded in response behavior, we firstly consider that the response process differs depending on the kind of survey question (Tourangeau, Rips, & Rasinski, 2000). Open-ended, semi-open-ended and closed-ended questions are used within survey questionnaires. In terms of open-ended questions, either text or numerical information is asked for. Closed-ended questions are divided into scalar questions, which use rating scales for a graduation of answers, and into questions with nominal lists of categories (check all that apply or check the most appropriate category). Another special kind of question is filter questions, which enable different pathways through the questionnaire.

Second, we should deliberate on the question of how falsifiers may differ from real respondents in terms of their motivation and response behavior. Motivation of falsifiers is made up of two different and, at the same time, conflicting aspects (Menold, Winker, Storfinger, & Kemper, 2013): On the one hand, falsifiers work carefully to avoid detection; on the other hand, falsifiers try to save time and reduce their effort whenever possible. Considering semi-open-ended questions, reduced interviewer effort can explain the results obtained by Bredl, Kötschau, and Winker (2012), who found that falsifiers avoid choosing the category "others, please specify." Reduced effort by falsifiers has also been found with filter questions (Bredl, Kötschau, & Winker, 2012; Hood & Bushery, 1997). For a health survey in the USA, Hood and Bushery (1997) report that falsifiers selected "the shortest path through the interview" producing survey participants who live in one person families and are "white, non-smoker, no health problems and no health insurance" (p. 821).

In the case of increased effort, falsifiers might respond to survey questions more carefully than real participants, showing less satisficing behavior. Satisficing means that survey respondents minimize their cognitive effort and produce satisfactory responses rather than optimal responses (Krosnick & Alwin, 1987). For example, this results in frequent selection of the option "do not know" or in providing a neutral position. Several results from previous research on falsifications within real survey data (which are not compatible with the explanation of reduced effort on the part of falsifiers) are in line with the assumption of reduced satisficing by falsifiers. This applies for less extreme responses to closed-ended scalar questions and less INR in both open-ended and closed-ended questions, which have been found for falsifiers (Bredl, Kötschau, & Winker, 2012; Schraepler & Wagner, 2005). Within closed-ended scalar questions reduced

satisficing by falsifiers can also be expected owing to other response tendencies, for example middle responding (providing neutral positions) and acquiescence, meaning providing "yes/agree" answers, regardless of item content (Messick, 1967). The next indicator that has been used in existing surveys (Murphy, Baxter, Eyerman, Cunningham, & Kennet, 2004; Porras & English, 2004; Schraepler & Wagner, 2005) is lower variance across responses reported for false data. This can be associated with the usage of a narrow range of response categories in rating scales (non-differentiation). A high non-differentiation in false data also can be associated with applying stereotypes during the falsifying process (Schnell, 1991).

For closed-ended questions with nominal response alternatives (but also for vertically ordered response alternatives in scalar questions), primacy and recency effects have been observed, and these are associated with the order in which response alternatives are presented (Tourangeau, Rips, & Rasinski, 2000). In the case of primacy effect, respondents prefer options appearing at the beginning of a list, whereas recency effect refers to a preference for categories placed at the end of a list. Both effects are influenced by presentation mode and both are associated with satisficing. Although primacy effect has been found for visually presented vertically listed response alternatives, recency effect is common in the case of acoustical presentation. In the case of visual presentation of response alternatives, falsifiers who are expected to show less satisficing should produce a lower primacy effect than real respondents. In the case of acoustical presentation of response categories, falsifiers use visual representation for falsifications so that a recency effect should be seen in real data but not in falsified data.

In the case of open-ended numerical questions, rounding occurs because respondents tend to provide round numbers (Tourangeau, Rasinski, Jobe, Smith, & Pratt, 1997). As discussed earlier in the text, rounding behavior limits the usage of Benford's law as an indicator of falsified data. However, rounding itself can be used as a formal indicator. In this regard, it can be expected that falsifiers who show less satisficing also demonstrate less rounding than real respondents.

To summarize, a number of formal indicators can be identified when considering the psychology of survey response (Tourangeau, Rips, & Rasinski, 2000). Some have been used for detection purposes in a real survey setting (INR, extreme responding, non-differentiation, filters and semi-open questions). Some (middle responding, acquiescent responding, rounding, primacy and recency effects) were first tested through our experimental research. Considering the results of previous research and our theoretical deliberations for the current study, we expect to observe reduced effort by falsifiers when using filters and semi-open-ended questions. Hence, we expect to see increased effort with other indicators, namely, INR, extreme and middle responding, acquiescence, primacy effect, recency effect, and

rounding. In addition, we expect to see a higher non-differentiation in false than in real data.

Content-Related Indicators

Content-related indicators can be classified through differentiation of question domains typically used in questionnaires. Thus, survey questions are typically divided into those about behavior and those about attitudes (Tourangeau, Rips, & Rasinski, 2000). Behavioral questions are about activities (e.g., political behavior) and habits (e.g., substance use), but also about dates, durations, and frequencies of certain events and behaviors. Answers to questions about attitudes are judgments about persons, objects, events, or issues. Sometimes, social survey research is also interested in participants' knowledge (e.g., about political issues). Knowledge questions make up the next group of content-related indicators. Previous research has not addressed the question of whether falsifiers are able to correctly forecast knowledge of real participants. The final group consists of personality questions, which are widely used in psychological research, but are rarely used in social survey research. For example, the International Social Survey Program (ISSP) has used a short version of the Big Five Inventory, the BFI-10 (Rammstedt & John, 2007). Personality questions ask for self-assessments by respondents and are related to individual characteristics, for example extroversion or introversion. In this way, they refer to information known only to individuals themselves, in contrast to attitude questions, which can also refer to commonly shared views. Thus, personality questions in particular may be difficult to falsify. The power of personality questions to differentiate between real and false data has not been addressed by previous research either.

Research Questions

On the basis of these classifications, we specify our research questions as follows: (1) What formal indicators are powerful in differentiating between real and false data? We will analyse indicators used in previous research as well as additional indicators, which we have defined earlier in the text. (2) Do formal indicators show different power in differentiating between real and false data as compared with content-related indicators (i.e., responses to attitude, behavior, knowledge, and personality questions)?

Procedure

Research Design Study 1

The first study was conducted in October–November 2010 using data from ALLBUS/German ISSP 2008 (http://www.gesis.org/unser-angebot/

daten-analysieren/umfragedaten/). ALLBUS interviews are conducted in a face-to-face mode, and ISSP data are collected by ALLBUS interviewers in CASI mode. ALLBUS/ISSP use a probability sample of German residents. We used the method first applied by Hippler (1979). A subsample of $N = 105$ interviews was drawn from the ALLBUS/ISSP data. We considered only cases for which verification was obtained by survey controls that the interview had been actually conducted. For this ALLBUS/ISSP subsample, false data (also $N = 105$ interviews) were produced by 16 instructed study participants (falsifiers). Falsifiers were provided with descriptions of survey participants including information about place of residence, age, sex, marital status, nationality, education, confession, number of children, and party preference. The descriptions contained characteristics that a potential falsifier could obtain through a short interview. The study participants were briefed on the purpose of the study, namely, improvement of methods to identify falsifications of survey data, their role as falsifiers and the procedure. The "falsifiers" were instructed to imagine conducting a face-to-face interview with the person described and to fill out the survey questionnaire as the described person would have done. Data produced by falsifiers were then compared with corresponding interviews from ALLBUS/ISSP. Some of the study participants ($n = 7$) were experienced interviewers who were employed in the telephone laboratory of a research institute (they had conducted between 60 and 3000 interviews). Another portion of the interviewers ($n = 9$) was made up of social sciences students employed by this research institute. The latter were not experienced interviewers. Six of the falsifiers were males and 10 were females. The average age was 33.5 years. The majority of falsifiers ($n = 13$) had higher education entrance certificates.

Research Design Study 2

In the second study, we did not use existing survey data, instead we collected data by applying selected ALLBUS/ISSP 2008 questions, as well as some additional questions, and this allowed for more focused data collection to construct different indicators. The real data were collected in a face-to-face survey on students from a German university ($N = 710$ real survey participants; age $Md = 24$ years, $IQR = 3$; 60% females). The study was conducted in the summer of 2011. The real study participants were recruited on the campus and interviewed by $N = 78$ students (age $Md = 24$ years, $IQR = 3$; 59% female), who were hired as interviewers. We provided payment of ~10 euros per hour to the interviewers. For real study participants, an incentive of 2 euros was provided. On average, each interviewer conducted 9.1 ($SD = 1$) interviews. Each interview was audio recorded and checked after study completion to ensure that interviews with survey respondents were not falsified.

In a second step, survey data were fabricated in the laboratory by the 78 interviewers hired for the study (710 false interviews were obtained). All of our falsifiers had comparable experience in conducting face-to-face interviews, and they were all familiar with the questionnaire as well as the possible answers by respondents. For falsifications in the laboratory, each interviewer received randomly assigned person descriptions for respondents interviewed by other interviewers. Person descriptions contained information about sex, age, subject of study, number of semesters enrolled, marital status, residence, living conditions, and country of origin. The briefing and instructions given to the falsifiers were the same as in the first study. Interviewers were paid 3 euros per interview on average for the task of falsifying. To increase the motivation of falsifiers to avoid detection, we announced a contest with payment of 100 euros for each of the three participants ("falsifiers") from the group who eluded our detection methods. Using this approach, we obtained a data set of $N = 710$ falsified interviews corresponding to each of the $N = 710$ real interviews.

Construction of Indicators

Formal indicators. In the first and second studies, we used the following formal indicators:

1. SEMI-OPEN: Frequency of selecting "others, please specify" in four semi-open-ended questions.
2. FILTER: Here, we used ALLBUS questions about the father's education and occupation when the respondents were 15 years old. These questions were preceded by a filter question asking what the father was doing at this point in time. Here, a response category allowed all subsequent questions about the respondents' father to be skipped: father not known. The frequency of choosing this category was used as indicator.
3. OPEN: Frequency of responses to open-ended questions (understanding of left and right for left-right scale, magazines read, sports one does) was used.
4. INR: Frequency of item non-response across all questions posed in the questionnaire was obtained (except open-ended questions and questions about respondents' fathers).
5. For Extreme Responding Style (ERS), we counted the most extreme responses on the ratings scale, for example "1" and "5" on a five-point rating scale, across all rating measures.
6. Middle Responding Style (MRS) represents frequency of endorsing the middle category in uneven rating scales.
7. Acquiescent Responding Style (ARS) was estimated based on BFI-10, which contains five pairs of items, which share substantive content but

differ in terms of positive and negative item wording. We counted the number of agreement responses, regardless of item direction.

8. Non-Differentiation (ND): We calculated the standard deviation of a respondent's responses across all items in multi-item sets and obtained a mean value for each case.
9. ROUNDING: We obtained the frequency of providing rounded answers to open-ended numerical questions. These were questions about minutes of watching television (round numbers here are 30, 60, 90, 120, etc.), body-mass information, and (in the first study) household's income. In the second study, students were asked about income from different sources using single questions: from parents, from foundations, from jobs, and total income.
10. PRIMACY: We counted how often the first two categories were chosen in questions with nominal lists of response categories, which were visually presented (four questions).
11. RECENCY. For two questions, interviewers read five to six categories in a rating scale to respondents (also in ALLBUS, 2008). We calculated how often the last category was chosen when answering these questions.

Content-related indicators. For these, we used responses to all questions for a domain (behavior, attitude, knowledge, and personality) that were included in the questionnaires (Table 1). There were single item measures as well as combined scale values, which were based on principal component analysis that we conducted using ALLBUS/ISSP 2008 data. For two behavior measures (past and future political participation, which were asked about with a nominal list of 13 alternatives), summative index values were calculated. A number between 1 and 13 was assigned to each alternative. As an index value, the sum of the numerical values for endorsed categories was obtained. For another behavioral indicator (NEWS; used in the second study), we asked what magazines one reads, and we provided a list of magazines as response options in which half were fictive magazines. As an indicator, we counted how often fictive magazines were chosen.

Knowledge indicators were frequencies of correct answers to knowledge questions in both studies (KNOW indicator). The familiarity ratings for non-existent words were asked for within the Vocabulary and Overclaiming Test (VOCT-T, Ziegler, Kemper, & Rammstedt, 2013; used in Study 2) (VOCT indicator). For personality measures, we built means across the items, which the authors of an inventory suppose measuring a single dimension. Thus, for BFI-10, we used five scale values.

Data Analysis

The data were analyzed with SPSS 20. We used binary logistic regressions to predict the probability of a case belonging either to the real (=0) or to the

Table 1
Description of Variables Included in Questionnaires in Study 1 and Study 2

Domain	Variables
Behaviors	*Both studies:* – frequency of watching television (one question); – political participation in the future and political participation in the past (both with nominal list of 13 activities, for example to contact popular politician, to take part on an protest action, to take part on elections); – intended party vote behavior (one question) *Study 1:* – taking part on religious services (one question); – frequency of different free time activities (11 items, which are grouped to four factors: sports and playing (PLAY, four items), reading and hearing music (CULTURE, three items), meeting with friends and relatives (FRIENDS, two items), watching television, shopping (CONSUM, two items); – being stressed or bored within the free time (STRESS, two items) *Study 2:* – reading behavior with nominal list of newspapers including also non-existent media (NEWS)
Attitudes	*Both studies:* – political resignation (anomia, four items) – acceptance of social inequality (three items); – political efficacy (three items); – trust in politicians (two items); – women's rights (six items); – left-right orientation (one question); – subjective social class (one question); – country economic situation (one question), – economic situation self (one question)

(continued)

Table 1
Continued

Domain	Variables	
	Study 1: – political trust (11 items); – evaluation of democracy (one question); – satisfying with democracy in Germany (one question); – satisfying with German government (one question); – political interest (one question); – fair share for themselves (one question); – trust in fellow men (one question), – right-wing nationalism (eight items); – national pride (two items); – satisfaction with own healthy (one question)	
Personality	*Both studies:* Big Five Inventory-10, BFI-10; Rammstedt & John, 2007), used to construct indicators for five scale scores: openness to experience (Big5_O), conscientiousness (Big5_C), extraversion (Big5_E), agreeableness (Big5_A), neuroticism (Big5_N)	
		Study 2: Self-enhancement and a self-efficacy scale (Beierlein, Kovaleva, Kemper, & Rammstedt, 2012).
Knowledge	*Study 1:* Nine questions on political knowledge (KNOW)	*Study 2:* – General knowledge test (KNOW); – Vocabulary and Overclaiming Test (VOCT-T, Ziegler, Kemper, & Rammstedt, 2013)

false (=1) data. Predictors (independent variables) were indicators grouped into five sets: Formal indicators, behavioral, attitude, knowledge, and personality variables. In a first step, we obtained the distinct predictive power of each set of indicators. In a second step, we gradually added each indicator set to Model 1 with formal indicators. In this way, we added behavioral variables to formal indicators (Model 2). For the next model (Model 3), attitude variables were added to the predictors of Model 2. In Model 4, knowledge measures were added to the predictors of Model 3, and then personality measures were added to Model 4 (Model 5). Finally, we tested the predictive power of content-related indicators, which were taken together (Model 6). All models were compared with respect to the increase in explained variance and in correctly predicting membership for false versus real data.

Before the analysis the prerequisites of binary logistic regression were tested. The multicollinearity diagnostic and exploration of residual distributions revealed acceptable results for all predictors in both studies. In the case of violated linearity, values of independent variables were grouped to ensure a linear relationship with the dependent variable (applies to ND and vote behavior). If the building of groups was ineffective, variables were excluded from the analysis (applies to "national pride", Study 1).

Results of Study 1

Formal Indicators

In Table 2 (Model 1), it can be seen that the filter question about respondents' fathers enables significant differentiation between false and real data. As indicated by the positive regression coefficient in Table 2, falsifiers answered this question with "father unknown" more often than real respondents. No significant effect can be seen for semi-open-ended questions. We expected lower frequencies in false data than in real data for a set of indicators (INR, ERS, MRS, ARS, PRIMACY, RECENCY and ROUNDING). Such results can be seen for ERS, ROUNDING and recency effect, with respective regression coefficients significantly and negatively related to the dependent variable, meaning that falsifiers selected less extreme categories, used fewer round numbers and showed lower recency effect than real respondents. Contrary to our expectations, falsifiers showed ARS and INR significantly more often than real respondents. No significant relationship with the dependent variable can be seen for middle responding, non-differentiation, primacy effect and answering of open-ended text questions.

Nevertheless, taken together, the formal indicators were powerful in correctly predicting membership for false versus real data. Seventy-six percent of cases were correctly classified using these indicators as predictors (see Table 2). This is associated with a 26% increase in correct predictions

Table 2
Results of Binary Regression Models in Study 1

	Indicators	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		B (SE)	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*
Formal	OPEN	0.01 (0.03)	1.01	0.01 (0.03)	1.01	−0.01 (0.03)	0.99	−0.02 (0.03)	0.98	−0.06 (0.04)	0.94		
	SEMI-OPEN	0.14 (0.74)	1.15	−0.06 (0.86)	0.94	0.33 (0.94)	1.39	0.45 (0.94)	1.57	−0.69 (1.24)	0.50		
	ROUNDING	−0.52** (0.18)	0.59	−0.34 (0.20)	0.71	−0.36 (0.23)	0.70	−0.37 (0.23)	0.69	−0.49 (0.28)	0.61		
	FILTER	1.38*** (0.42)	3.97	1.33** (0.47)	3.78	1.57** (0.56)	4.81	1.56** (0.56)	4.75	2.07** (0.76)	7.90		
	INR	0.59* (0.27)	1.8	0.69* (0.30)	2.03	0.88* (0.39)	2.40	1.12** (0.43)	3.08	1.19* (0.61)	3.29		
	ERS	−0.07*** (0.02)	0.93	−0.06** (0.02)	0.94	−0.05 (0.03)	0.95	−0.05 (0.03)	0.96	−0.06 (0.03)	0.94		
	MRS	−0.02 (0.04)	0.98	−0.006 (0.04)	0.99	0.02 (0.06)	1.02	0.02 (0.06)	1.02	0.07 (0.07)	1.07		
	ARS	0.27** (0.1)	1.31	0.34** (0.11)	1.29	0.43** (0.14)	1.54	0.44** (0.15)	1.54	0.36 (0.19)	1.44		
	ND	0.33 (0.26)	1.39	0.45 (0.31)	1.56	0.48 (0.36)	1.63	0.39 (0.37)	1.47	0.38 (0.50)	1.47		
	RECENCY	−1.27*** (0.34)	0.28	−1.23*** (0.35)	0.29	−1.56*** (0.41)	0.21	−1.69*** (0.42)	0.19	−2.18*** (0.56)	0.11		
	PRIMACY	0.16 (0.12)	1.17	0.26 (0.14)	1.29	0.16 (0.17)	1.18	0.15 (0.17)	1.16	0.29 (0.23)	1.34		
Behavior	Past political participation			−0.02 (0.02)	0.98	−0.02 (0.02)	0.98	−0.02 (0.02)	0.98	−0.02 (0.02)	0.99	−0.03* (0.02)	0.97
	FRIENDS			−0.81** (0.28)	0.45	−1.10** (0.34)	0.33	−1.22*** (0.34)	0.30	−1.83** (0.54)	0.16	−0.97** (0.31)	0.38
	CONSUM			0.91** (0.33)	2.48	0.95** (0.36)	2.58	0.88* (0.36)	2.40	1.22* (0.51)	3.38	0.61* (0.30)	1.84
	STRESS			−0.44 (0.26)	0.65	−0.60(0.31)	0.55	−0.74* (0.33)	0.48	−1.14** (0.44)	0.32	−0.94*** (0.26)	0.39
Attitude	Political efficacy					1.01** (0.30)	2.75	0.91** (0.30)	2.47	1.42** (0.44)	4.16	1.04*** (0.27)	2.83
	Trust politicians					0.55* (0.25)	1.73	0.52* (0.25)	1.69	0.96** (0.33)	2.62	0.86*** (0.22)	2.37
	Womens' rights					−1.09** (0.34)	0.34	−1.15** (0.34)	0.32	−0.62 (0.44)	0.54	−0.53 (0.27)	0.59
	Nationalism					−0.22 (0.33)	0.80	−0.16 (0.33)	0.85	−0.46 (0.44)	0.63	−0.12 (0.27)	0.89
Knowledge	KNOW							0.16 (0.10)	1.18	0.21 (0.12)	1.24	0.02 (0.07)	1.02
Personality	Big5_A									−0.27 (0.38)	0.97	−0.20 (0.27)	1.22
	Big5_O									−1.77*** (0.45)	0.17	−1.11*** (0.27)	0.33
	Big5_N									−0.19 (0.37)	0.83	−0.21 (0.24)	1.23
	Big5_E									1.07** (0.39)	2.90	0.63** (0.23)	1.88
	Big5_C									0.20 (0.41)	1.22	0.51 (0.27)	1.67
	$R^2_{(Cox\ \&\ Snell)}$	.34		.40		.49		.50		.57		0.38	
	$R^2_{(Nagelkerke)}$	.46		.53		.66		.67		.76		0.51	
	Model $\chi^2_{(df)}$	$88.07_{(11)}$***		$106.79_{(15)}$***		$141.59_{(19)}$***		$144.24_{(20)}$***		$163.918_{(25)}$***		$93.77_{(14)}$***	
	n hit/n real	81/105		84/104		90/104		89/104		93/104		86/104	
	n hit/n false	79/105		79/105		86/103		84/103		78/90		65/90	
	% hit rate	76%		78%		85%		84%		88%		78%	

Note. ***$p < .001$; **$p < .01$; *$p < .05$. Dependent variable falsified data (yes = 1; no = 0); For explanation of predictors in Model 1, see text, p. 9–10; for remaining predictors see text and Table 1. Results for models with equal sample sizes ([$N = 194$] $R^2_{(Nagelkerke)}$): Model 1 (.44), Model 2 (.52), Model 3 (.66), Model 4 (.66).

when compared with the baseline model. With the latter, all cases are classified into a group with 50% correct predictions. The model with formal indicators also has high variance, which is explained in the dependent variable.

Content-Related Indicators

For the individual model test including only behavior measures (10 variables, see Table 1), we obtained a satisfactory model fit ($\chi^2_{(10,\ N=208)} = 36.51$, $p < .001$) with a lower explained variance ($R^2_{(Nagelkerke)} = .22$) and a lower prediction accuracy (69%) than for the model with only formal indicators. A significant relationship with the dependent variable was obtained for only 4 of 10 predictors: past political participation, meeting friends (FRIENDS), watching TV and shopping (CONSUM) and being stressed during free time (STRESS). Eighteen variables were used to analyze a distinct effect of attitude indicators (Table 1). The results show that attitude variables had a significant impact on the dependent variable ($\chi^2_{(19,\ N=194)} = 65.12$, $p < .001$). The model leads to $R^2_{(\text{Nagelkerke})} = 0.38$ explained variance, and the accuracy of prediction is ~75%. However, significant regression coefficients were found for only four of 18 variables (political efficacy, political trust, womens' rights, right-wing nationalism). Using the political knowledge indicator as a single predictor did not reveal a significant effect ($\chi^2_{(1,\ N=210)} = 0.07$, $p > .10$). If personality variables are chosen for a distinct prediction, they lead to a significant model, but with low predictive power ($\chi^2_{(5,\ N=197)} = 18.17$, $p < .01$; $R^2_{(\text{Nagelkerke})} = .12$; 60% correct predictions).

The successive addition of all behavioral and attitude measures to formal indicators was not possible in Study 1, as this resulted in too many predictors for the given sample size. To avoid this problem, we selected variables that showed a significant relationship with the dependent variable within the distinct analyses conducted for each indicator group (see earlier in the text). We added four behavioral variables (past political activities, FRIENDS, CONSUM, and STRESS) to formal indicators. The results are presented in Table 2 (Model 2). Despite an increase in explained variance, adding the four behavioral variables did not remarkably improve the predictive accuracy of Model 1. Within Model 2, the regression coefficients for past political activities and STRESS lost significance. As compared with Model 1, rounding also lost significance. Adding four attitude variables (those with significant regression coefficients obtained from the separate regression analysis, see earlier in the text) to the predictors of Model 2 increased the explained variance and the predictive model power (Model 3, Table 2). However, right-wing nationalism did not have a significant regression coefficient within this model. In the next model (Model 4), we added the knowledge measure (KNOW) to the predictors included in Model 3, but this did not change the results.

In Model 5, personality variables were added to Model 4. Model 5 showed an increase in explained variance and predictive accuracy (now ~88%).

Model 6 included all content-related indicators from Models 2 to 5, whereas formal indicators were excluded. The results with regard to explained variance and correct predictions were comparable with those of the model with formal indicators (Model 1). However, like in Model 5, the cost was a remarkable decrease in the number of cases included in the analysis. Using data from Models 5 and 6 ($N = 194$) for other models did not remarkably change the regression coefficients or the model fit (Nagelkerke R^2 values for these models are reported in the note for Table 2). Thus, Models 1 to 6 can be compared using the explained variances.

Results of Study 2

Formal Indicators

Formal indicators were powerful in explaining differences between false and real data. Model 1 (Table 3) enabled correct membership prediction for real versus false data 68% of the time and explained 22% of variance in the dependent variable. The regression coefficients of nearly all formal indicators were significant. Non-significant coefficients can be seen for semi-open-ended questions and INR. Significant results show that falsifiers used the option "father not known" as a filter significantly more often than real respondents (FILTER). Next, the false data exhibited significantly lower primacy and recency effects, as well as significantly higher ND, than real data. For other indicators, we can see less MRS and less ARS in the false data. In addition, lower frequency of responding to open-ended questions (OPEN), higher rounding, and higher ERS were found in the false data than in real data.

Content-Related Indicators

In contrast to the first study, we used only a few behavioral indicators in the second study (five variables, see Table 1). Nevertheless, these variables demonstrated a high level of power in differentiating between real and false data. However, the variable "watching television", which was not significantly correlated with the dependent variable, produced a remarkable number of missing values. We excluded this variable from further analyses, as this did not significantly change the results. With the remaining four variables, the model fit is described as follows: $\chi^2_{(4,\ N=1415)} = 165.45$, $p < .001$; $R^2_{(Nagelkerke)} = .15$; 65% of correct predictions. This high effect is comparable with the single effect of formal indicators. Regression coefficients are significant for all four behavioral variables and show that falsifiers endorsed fictitious magazines more often in describing respondents' reading behavior (NEWS),

Table 3
Results of Binary Regression Models in Study 2

	Indicators	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		B (SE)	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*
Formal	INR	−0.01 (0.03)	0.99	−0.01 (0.03)	0.98	−0.05 (0.03)	0.95	−0.07* (0.03)	0.93	−0.08* (0.04)	0.92		
	SEMI-OPEN	−0.30 (0.17)	0.72	−0.83*** (0.21)	0.44	−1.05*** (0.22)	0.35	−1.03*** (0.23)	0.36	−1.07*** (0.23)	0.34		
	ROUNDING	1.99*** (0.27)	7.35	1.85*** (0.29)	6.38	1.68*** (0.3)	5.36	1.71*** (0.31)	5.51	1.66*** (0.31)	5.28		
	FILTER	0.94*** (0.2)	2.57	1.08*** (0.21)	2.96	1.01*** (0.22)	2.75	1.02*** (0.22)	2.76	1.02*** (0.23)	2.78		
	OPEN	−0.32*** (0.09)	0.73	−0.29** (0.1)	0.74	−0.48*** (0.12)	0.62	−0.4*** (0.12)	0.61	−0.48*** (0.13)	0.62		
	ERS	0.07*** (0.01)	1.07	0.09*** (0.01)	1.1	0.07*** (0.02)	1.08	0.07*** (0.02)	1.08	0.04* (0.02)	1.04		
	MRS	−0.05* (0.02)	0.95	−0.04 (0.02)	0.96	−0.04 (0.03)	0.96	−0.04 (0.03)	0.96	0.02 (0.03)	1.02		
	ARS	−0.45*** (0.06)	0.64	−0.48*** (0.06)	0.62	−0.54*** (0.07)	0.58	−0.52*** (0.07)	0.59	−0.5*** (0.08)	0.6		
	ND	0.91*** (0.1)	2.48	1.0*** (0.1)	2.73	0.82*** (0.11)	2.28	0.86*** (0.12)	2.37	0.7*** (0.12)	2.01		
	RECENCY	−0.53* (0.21)	0.60	−0.39 (0.22)	0.68	−0.51* (0.24)	0.6	−0.54* (0.24)	0.58	−0.46 (0.24)	0.63		
	PRIMACY	−0.33*** (0.11)	0.72	−0.21 (0.12)	0.81	−0.26* (0.12)	0.77	−0.25* (0.12)	0.78	−0.24 (0.13)	0.78		
Behaviour	NEWS			1.32*** (0.32)	3.76	1.23*** (0.33)	3.43	1.11*** (0.33)	3.04	1.05*** (0.33)	2.87	0.93** (3.22)	2.53
	Future political participation			0.04*** (0.00)	1.04	0.05*** (0.00)	1.05	0.05*** (0.00)	1.05	0.05*** (0.00)	1.05	0.05*** (0.01)	1.05
	Past. political participation			−0.05*** (0.01)	0.95	−0.05*** (0.01)	0.95	−0.05*** (0.01)	0.95	−0.05*** (0.01)	0.95	−0.06*** (0.01)	0.95
	Party vote			−0.02*** (0.00)	0.98	−0.03*** (0.00)	0.97	−0.03*** (0.00)	0.97	−0.03*** (0.00)	0.97	−0.01** (0.00)	0.99
Attitude	Anomia					0.6** (0.23)	1.82	0.67** (0.24)	1.96	0.54* (0.25)	1.71	0.63** (0.22)	1.88
	Social Equity					−0.16 (0.12)	0.85	−0.17 (0.12)	0.84	−0.28* (0.12)	0.75	−0.38*** (0.11)	0.68
	Political efficacy					−0.11 (0.06)	0.9	−0.11 (0.06)	0.89	−0.07 (0.06)	0.93	0.05 (0.06)	1.05
	Political trust					0.3*** (0.07)	1.35	0.32*** (0.07)	1.37	0.31*** (0.08)	1.36	0.27*** (0.07)	1.31
	Womens' rights					0.52** (0.18)	1.68	0.46* (0.18)	1.58	0.36 (0.19)	1.43	0.41* (0.17)	1.50
	Left-right orientation					0.02** (0.01)	1.02	0.02** (0.01)	1.02	0.02* (0.01)	1.02	0.02** (0.01)	1.02
	Social layer					0.14 (0.1)	1.15	0.15 (0.1)	1.16	0.06 (0.11)	1.07	0.02 (0.10)	1.02
	Oec. situation country					−0.33* (0.14)	0.72	−0.37** (0.14)	0.69	−0.46** (0.14)	0.63	−0.45*** (0.13)	0.64
	Oec. situation self					0.01 (0.14)	1.01	0.03 (0.14)	1.03	0.05 (0.14)	1.06	0.18 (0.13)	1.20

(continued)

Table 3
Continued

	Indicators	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
		B (SE)	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*	*B (SE)*	*Exp (B)*
Knowledge	VOCT							−0.83*** (0.21)	0.43	−0.85*** (0.21)	0.43	−0.74*** (0.19)	0.48
	KNOW							−0.04 (0.05)	0.96	−0.07 (0.05)	0.93	−0.06 (0.04)	0.94
Personality	Big5_A									−0.05 (0.09)	0.95	−0.28** (0.08)	0.76
	Big5_O									−0.17 (0.08)	0.85	−0.23** (0.08)	0.78
	Big5_N									−0.00 (0.08)	0.99	−0.13 (0.07)	0.87
	Big5_E									−0.03 (0.11)	0.97	−0.21* (0.10)	0.82
	Big5_C									0.36*** (0.1)	1.43	0.22** (0.09)	1.25
	Self-enhancement									0.06 (0.1)	1.06	0.01 (0.09)	1.01
	Self-efficacy									0.25*** (0.07)	1.29	0.24*** (0.05)	1.27
	R^2 *(Cox & Snell)*	.17		.25		.28		.29		.3		0.19	
	R^2 *(Nagelkerke)*	.22		.34		.37		.38		.41		0.25	
	Model χ^2 *(df)*	260.40 (11)***		414.46 (15)***		452.07 (24)***		470.07 (26)***		500.74 (33)***		288.74 (22)***	
	n hit/n real	486/710		525/709		522/698		524/698		524/696		490/696	
	n hit/n false	473/710		511/706		510/687		517/687		515/683		450/683	
	% hit rate	68%		73%		75%		75%		75%		68%	

Note. ***$p < .001$; **$p < .01$; *$p < .05$. Dependent variable falsified data (yes = 1; no = 0); For explanation of predictors in Model 1, see text, p. 9–10; for remaining predictors, see text and Table 1. Results for models with equal sample sizes [$N = 1379$] R^2 *(Nagelkerke)*: Model 1 (.22), Model 2 (.34), Model 3 (.37), Model 4 (.38).

they overestimated participants' future political participation, and they underestimated real respondents' past political participation. In addition, falsifiers were not able to correctly display participants' intended voting behavior. In contrast to behavioral variables, the distinct impact of nine attitude variables on the dependent variable was low ($\chi^2_{(5,\ N=1390)} = 43.76$, $p < .001$; $R^2_{(Nagelkerke)} = 0.04$; 57% correct predictions). However, there were significant relationships between four attitude variables (anomia, political trust, womens' rights, left-right orientation, and satisfaction with countries economic situation, see also Table 3) and the dependent variable. If the two knowledge variables were used within a model to obtain their distinct effect, it was also weak ($\chi^2_{(2,\ N=1420)} = 24.54$, $p < .001$; $R^2_{(Nagelkerke)} = 0.02$; 57% correct predictions). However, it was found that falsifiers used non-existent terms significantly more often than respondents when describing vocabulary knowledge (VOCT). The model that included only personality variables also showed a small effect on the dependent variable ($\chi^2_{(7,\ N=1414)} = 42.82$, $p < .001$; $R^2_{(Nagelkerke)} = .04$; 57% correct predictions).

Adding behavioral variables to the formal indicators remarkably improved the model fit (Model 2 in Table 3), increasing Nagelkerke $R^2 = .22$ to a value of $R^2 = .34$ and the proportion of correct predictions from 68 to 73%. Adding attitude variables to Model 2 did not remarkably improve its predictive power (Model 3): the number of correct predictions increased by only 2%, and at the same time, this reduced the sample size, as there was a number of missing values within the attitude variables. As with attitude variables, we were not able to obtain better model prediction if knowledge variables were added to Model 3 (Model 4 in Table 3). Adding personality variables to the predictors in Model 4 (Model 5) did not remarkably increase its predictive power either. However, Model 6, which included all content-related indicators but not formal indicators, produced explained variances and hit rates comparable with those of Model 1. Using data from Model 5/6 with N = 1379 for the other four models did not remarkably change the results (Nagelkerke R^2 values for these models are reported in the note of Table 3).

Now, we will look at changes in formal indicators when content-related indicators were added to Model 1. We can see that the impact of INR becomes significant across Models 4 and 5. Here, a lower INR can be seen in the false data. The effect of semi-open-ended questions becomes significant in Models 2–5. However, primacy and recency effects lost their significant impact in Model 5.

Comparison of the Results of Study 1 and Study 2

Within both studies, we found that formal as well as content-related indicators were powerful in predicting membership as false versus real data. However, the power of content-related indicators varied between the two studies. Thus,

a strong relationship with real versus false data was found for behavioral variables in Study 2, whereas in Study 1 this relationship was rather weak. In contrast to the first study, in the second study we did not provide the falsifiers with information about party preference of real participants. This may explain the large differences in responses to questions about political behavior between real and false data in the second study. Some attitude variables showed substantial improvement of predictive power of indicators in the first study, but not in the second study. However, that was the case for only a small set out of the numerous attitude variables we used. The result, when considering behavioral variables, supports the findings by Hippler (1979), who showed that the accuracy of predicting real respondents' responses depends on what information falsifiers know about respondents. Knowledge and personality variables only marginally enabled differentiation between real and false data. However, several variables, for example the VOCT indicator based on fictive response alternatives, enabled significant level of distinction between real and false data.

Concerning formal indicators, we expected to obtain more frequent usage of filter options by falsifiers as well as less usage of "others" in semi-open-ended questions (see e.g., Bredl, Kötschau, & Winker, 2013, for a real survey setting). Both of these effects can be explained by reduced effort during the falsification process. Concerning filter option, the results of both studies are in line with our expectations, whereas a significant result for semi-open-ended questions was obtained in the second study only. Similarly, the result for non-differentiation was in line with our expectation (and with results by, for example, Schraepler & Wagner, 2005) in the second study only.

For other formal indicators, we expected to obtain differences between real and false data, which can be explained by increased effort on the part of falsifiers attempting to show less satisficing to produce "proper" data. Consistent results across the two studies were obtained for recency effect, which—as expected—was lower in the false data. The results of the two studies are mixed with regard to other indicators. In the first study, results for rounding and extreme responding correspond with the expectation of increased effort on the part of falsifiers. In the second study, results concerning acquiescence, INR, middle responding and primacy effect are in line with this explanation. For ERS, ARS and rounding, we found different directions in correlations with the dependent variable in the two studies.

An explanation for the mixed results may be that falsifiers differed between the two studies. In the first study, a portion of falsifiers was made up of inexperienced interviewers. In the second study, the falsifier group was homogeneous (all were students with comparable experience in conducting interviews on the given topic). With the data from the first study, Menold and Opper (2013) found that experienced interviewers differed more than inexperienced interviewers in values of several formal indicators from the real data.

The different relationships of ARS and ERS with false data in two studies may be explained by multivariate interactions between several formal indicators. ARS was positively associated with false data in the first study, but this relationship was negative in the second study. In contrast, ERS was negatively associated with false data in the first study, but this relationship was positive in the second study. Using additional analyses, we tested the assumption regarding an impact of multivariate interactions on these relations. With the data of the first study, we included one-way and two-way interactions of ARS with ERS, INR, and ND in the model with formal indicators (Model 1, Table 2). In this way, we considered other significant formal indicators, which also used information from multi-item sets. We found the one-way interaction between ARS and ND and all two-way interactions to be significant ($p < .05$). Including that interactions in Model 1 (first study) lead to a new negative (!) significant regression coefficient for ARS ($B = -1.92$; $SE = 0.85$, $p < .05$). In this way, the correlation direction of ARS with the independent variable changed to the direction, which is comparable with the result of the second study as well as with the explanation of increased effort on the part of falsifiers. In the second study, a change in the regression coefficient of ERS to $B = -.07$ ($SE = 0.03$; $p < .01$) was visible, if an interaction term between ERS and ND was included in Model 1 (Table 3). This regression coefficient was comparable with the one found in the first study, and it was in line with the expectation of lower satisficing on the part of falsifiers. Therefore, interactions between formal indicators were responsible for the mixed results concerning ARS and ERS. Taking these interactions into account allowed for consistent results across both studies, which are also in line with the expectation of less falsifiers' satisficing.

Another explanation for mixed results could be differences in the questionnaires used in the two studies. For example, in the case of rounding as an indicator, questions about income varied between the two studies. It was not appropriate to ask the students about their total household net income, although this was asked about in the ALLBUS (and in Study 1). Instead, incomes from different sources—parents and relatives, foundations, jobs—were asked about in the second study. The classic total net income question commonly used in surveys poses memory problems for real respondents, resulting in high rounding. In contrast, in the case of specific single questions used in the second study, students knew the requested information and reported it exactly, whereas falsifiers had problems providing exact information. Thus, falsifiers produced more round numerical answers than real respondents.

Finally, in the two studies, falsifiers differed concerning their motivation to avoid detection. In the second study, falsifiers were paid for their work, whereas in the first study, no direct payment was provided. In addition, in the second study, we provided a reward for the best falsifiers, meaning those who could not be detected by our analytical methods.

Summary and Conclusion

The aim of the studies presented was to reveal differences between real and false data that can be used as indicators to identify false data. We classified indicators into two groups: Formal indicators, which are based on differences in response behavior between real respondents and falsifiers, and content-related indicators, which are based on differences in the content of responses.

Our first research question addressed the power of each of the formal indicators to differentiate between real and false data. In the second study in particular, we found that all formal indicators were powerful in this regard. Concerning indicators, used in previous research, we obtained results that were comparable with those from previous studies (Bredl, Kötschau, & Winker, 2012; Schraepler & Wagner, 2005). This applies for filter questions in both studies, extreme responses in the first study, and semi-open questions and non-differentiation in the second study. Mixed results between the two studies could be explained by multivariate interactions, differences in questionnaires as well as interviewers' experience and their general effort to avoid detection. Thus, falsifiers in the second study were homogeneous in terms of their professional interviewers' experience, and they appeared to be more motivated to avoid detection, based on the payment methods used. At the same time, they produced results that are compatible with the results obtained in a real survey context and those that are in line with our expectations. Therefore, increased effort by falsifiers and better visibility regarding this in the data through associated formal indicators could also be expected in a real survey context, in which dishonest interviewers are of course strongly interested in avoiding detection.

Regarding the second research question, we obtained clear and comparable results from both studies. These show that formal indicators are powerful in predicting the probability of data being false or real. Content-related indicators also show a significant amount of power in differentiating between real and false data. However, in the first study, significant differences between real and false data were obtained only for a small set of a large number of content-related indicators. Next, some content-related indicators produced a considerable number of missing values and their performance seemed to depend on the information available to interviewers about respondents (Hippler, 1979). This restricts the usability of content-related indicators for detection purposes. Nevertheless, content-related indicators, such as reading newspapers (NEWS) and vocabulary knowledge (VOCT), can be combined with formal indicators owing to their high level of effectiveness in differentiating between real and false data. NEWS and VOCT use fictitious information included in response alternatives, which appears to be useful for creating indicators for detection purposes. Menold, Winker, Storfinger, and Kemper (2013) used formal indicators from the second study, combining them with past political

behavior, NEWS and VOCT to detect false data by means of a multivariate cluster analysis. They obtained promising results with almost 90% correct differentiation between real and false data. Usage of multivariate detection methods may also help to avoid problems resulting from multivariate interactions between formal indicators.

Our results, which are comparable with those obtained in real survey contexts, indicate that the behavior of our falsifiers was similar to that shown in real surveys. However, more studies are needed in a real survey setting to verify our experimental results in the context of real surveys, in particular as many indicators were addressed for the first time in our research. In addition, there is a need for more research on the effects of interviewer experience and interviewer motivation to avoid detection on the power of different indicators to distinguish between real and false data. Finally, research on the motivation of interviewers to not deviate should be conducted, and it should address such factors as interview length, interviewer payment, work pressure (e.g., requiring high response rates), control procedures (e.g., Crespi, 1945), practices to foster interviewers' commitment, or an organizational culture, which acknowledges interviewers' work (Gwartney, 2013).

Funding

This work was supported by the German Research Foundation (DFG) [WI 2024/2-1 and ME 3538/2-1].

Acknowledgements

We would like to thank Peter Winker, Nina Storfinger, and Viktoria Trofimow for their helpful support in the research presented in this article.

References

AAPOR. *Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects.* Retrieved from http://www.aapor.org/Content/aapor/AdvocacyandInitiatives/StandardsandEthics/InterviewerFalsificationPracticesandPolicies/ReporttoAAPORStandardsCommonInterviewerFalsification/default.htm.

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU), *GESIS Working Papers 17*. Retrieved from http://www.gesis.org/publikationen/gesis-working-papers/.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78*(1), 551–572.

Biemer, P. P., & Stokes, S. L. (1989). The optimal design of quality control samples to detect interviewer cheating. *Journal of Official Statistics, 5*(1), 23–39.

Bredl, S., Storfinger, N., & Menold, N. (2013). A literature review of methods to detect fabricated survey data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys* (pp. 3–24). Frankfurt: Peter Lang Academic Research.

Bredl, S., Kötschau, K., & Winker, P. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology, 38*, 1–10. Retrieved 2013, from http://publications.gc.ca/collections/collection_2012/statcan/12-001-x/12-001-x2012001-eng.pdf.

Bushery, J. M. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the Survey Research Methods Section, ASA,* (9), 316–320.

Case, P. B. (1971). How to catch interviewer errors. *Journal of Advertising Research, 11*(2), 39–43.

Crespi, L. P. (1945). The Cheater Problem in Polling. *Public Opinion Quarterly, 9*(4), 431–445.

Evans, F. B. (1961). On interviewer cheating. *Public Opinion Quarterly, 25*(1), 126–127.

Gwartney, P. A. (2013). Mitschief versus mistakes: motivating interviewers to not deviate. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys* (pp. 195–215). Frankfurt: Peter Lang Academic Research.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New Jersey: Wiley.

Hippler, H. J. (1979). Untersuchung zur "Qualität" absichtlich gefälschter Interviews. *ZUMA-Arbeitspapier,* 1–30.

Hood, C. C., & Bushery, J. M. (1997). Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *Proceedings of the Survey Research Methods Section, ASA,* (27), 820–824.

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

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201–219.

Menold, N., Winker, P., Storfinger, N., & Kemper, C. J. (2013). A method for ex-post identification of falsifications in survey data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys* (pp. 25–48). Frankfurt: Peter Lang Academic Research.

Menold, N., & Opper, M. K. (2013). Interviewer experience and "quality" of falsified data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' deviations in surveys* (pp. 161–172). Frankfurt: Peter Lang Academic Research.

Messick, S. J. (1967). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 115–145). Chicago: Aldine.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D., & Kennet, J. (2004). A system for detecting interviewer falsification. *American Association for Public Opinion Research 59th Annual Conference,* 4968–4975.

Porras, J., & English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. *Proceedings of the American Statistical Association (Survey Research Methods Section),* 4223–4228.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. doi:10.1016/j.jrp.2006.02.001.

Reuband, K. H. (1990). Interviews, die keine sind: "Erfolge" und "Mißerfolge" beim Fälschen von Interviews. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42*(4), 706–733.

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

Schraepler, 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.*

Schraepler, J.-P., & Wagner, G. (2005). Characteristics and impact of faked interviews in surveys – an analysis of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv, 89*, 7–20.

Tourangeau, R., Rasinski, K. A., Jobe, J. B., Smith, T. W., & Pratt, W. F. (1997). Sources of error in a survey on sexual behavior. *Journal of Official Statistics, 13*(4), 341–365.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Turner, C., Gribble, J., Al-Tayyip, A., & Chromy, J. (2002). Falsification in epidemiologic surveys: Detection and remediation (Prepublication Draft). *Technical Papers on Health and Behavior Measurement, No. 53*. Washington DC: Research Triangle Institute.

Wang, Y., & Pedlow, S. (2005). Detecting Falsified Cases in SCF 2004 Using Benford's Law. *Proceedings of the American Statistical Association (Survey Research Methods Section),* 3652–3657.

Ziegler, M., Kemper, C. J., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). *Journal of Individual Differences, 34*(1), 32–40. doi: 10.1027/1614-0001/a000093.

Biographical Notes

Dr. Natalja Menold completed a Master's degree in psychology in 2000 and received her doctorate in 2006. Since 2007, she has been working at GESIS in Mannheim, Germany. Her current position is senior survey methodologist and researcher, head of the team "survey instrumentation."

Dr. Christoph Kemper is a psychologist and senior researcher. In his research, he focuses on the assessment of normal and abnormal personality, cognitive assessment as well as different aspects of data quality and related interindividual differences, for example response styles.