Secondary Publication

Olbrich, Lukas; Kosyakova, Yuliya; Sakshaug, Joseph W.; Schwanhäuser, Silvia

Detecting Interviewer Fraud Using Multilevel Models

Date of secondary publication: 03.05.2024

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-950437

Primary publication

Olbrich, Lukas; Kosyakova, Yuliya; Sakshaug, Joseph W.; Schwanhäuser, Siliva (2024): „Detecting Interviewer Fraud Using Multilevel Models“. In: Journal of survey statistics and methodology : JSSAM, Vol. 12, Nr. 1, pp. 14-35, Oxford: Oxford Univ. Press, doi: 10.1093/jssam/smac036.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:

https://creativecommons.org/licenses/by/4.0/legalcode

DETECTING INTERVIEWER FRAUD USING MULTILEVEL MODELS

LUKAS OLBRICH*
YULIYA KOSYAKOVA
JOSEPH W. SAKSHAUG
SILVIA SCHWANHÄUSER

Interviewer falsification, such as the complete or partial fabrication of interview data, has been shown to substantially affect the results of survey data. In this study, we apply a method to identify falsifying face-to-face interviewers based on the development of their behavior over the survey field period. We postulate four potential falsifier types: steady low-effort falsifiers, steady high-effort falsifiers, learning falsifiers, and sudden falsifiers. Using large-scale survey data from Germany with

Lukas Olbrich is a PhD candidate in the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany and the Ludwig-Maximilian University of Munich, Germany. Yuliya Kosyakova is a post-doctoral researcher at the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany and an associate lecturer at the Otto-Friedrich University of Bamberg, Germany. Joseph W. Sakshaug is Professor of Statistics at the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany, the Ludwig-Maximilian University of Munich, Germany, and the University of Mannheim, Germany. Silvia Schwanhäuser is a PhD candidate in the Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany and the University of Mannheim, Germany.

The authors thank Regina Riphahn, Daniel Kühnle, and Marieke Volkert for engaging in insightful discussions at various stages of the research process and all those who commented on the earlier presentations of this work at the following events: the 8th Conference of the European Survey Research Association (ESRA) in Zagreb (July 2019) and the 75th Annual Conference of the American Association for Public Opinion Research (virtual). The authors also thank the Editor, the Associate Editor, and three anonymous referees for helpful comments and suggestions. The authors also acknowledge the financial support of the Charles Cannell Fund in Survey Methodology awarded to Yuliya Kosyakova. Lukas Olbrich gratefully acknowledges financial support from the Graduate Programme of the Institute for Employment Research (IAB) and the University of Erlangen-Nuremberg (GradAB). Data access was provided via researcher contacts at the Institute for Employment Research (IAB). External researchers may apply for access to these data by submitting a user-contract application to the SOEP Research Data Center (https://www.diw.de/en/diw_02.c.222836.en/data_access_and_order.html). Code for replication is available at https://osf.io/r8y9p/. The study design and analysis were not preregistered.

*Address correspondence to Lukas Olbrich, Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany; E-mail: lukas.olbrich@iab.de.

https://doi.org/10.1093/jssam/smac036 Advance access publication 2 January 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the American Association for Public Opinion Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

verified falsifications, we apply multilevel models with interviewer effects on the intercept, scale, and slope of the interview sequence to test whether falsifiers can be detected based on their dynamic behavior. In addition to identifying a rather high-effort falsifier previously detected by the survey organization, the model flagged two additional suspicious interviewers exhibiting learning behavior, who were subsequently classified as deviant by the survey organization. We additionally apply the analysis approach to publicly available cross-national survey data and find multiple interviewers who show behavior consistent with the postulated falsifier types.

KEYWORDS: Interviewer behavior; Interviewer effects; Interviewer falsification; Multilevel modeling.

Statement of Significance

This study proposes a new method to identify fraudulent interviewers in face-to-face surveys. In particular, we investigate whether falsifying interviewers can be identified by their dynamic behavior over the field period. We postulate four falsifier types: steady low-effort falsifiers, steady high-effort falsifiers, learning falsifiers, and sudden falsifiers. These falsifier types are tested using complex multilevel models applied to German survey data containing verified cases of interviewer falsification. Focusing on the behavior over the field period allows for identifying a verified falsifier and two previously undetected fraudulent interviewers. Applying these methods to further publicly available survey data, we also find behavior expected of the postulated falsifier types. Our findings show that fraudulent interviewers can use sophisticated strategies to avoid detection.

1. INTRODUCTION

Interviewers are a well-known error source in face-to-face surveys. Researchers have intensively investigated unintentional interviewer errors such as accidentally skipping questions or recording responses in error (Weisberg 2005). Less is known about interviewer falsification—defined by the American Association for Public Opinion Research (AAPOR) as “the intentional departure from the designed interviewer guidelines or instructions, unreported by the interviewer, which could result in the contamination of data” (AAPOR 2003, p. 1). Interviewer falsification can take many forms, from strategically miscoding responses to avoid follow-up questions to falsifying complete or partial interviews (AAPOR 2003). In this study, we focus on the falsification of interviews.

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

Given a lack of publicly available data on verified falsifications, evidence on the prevalence and extent of interviewer falsification is rare. While the reported share of completely falsified interviews rarely exceeds 5 percent in large-scale surveys (Bredl et al. 2013; Finn and Ranchhod 2017; Robbins 2019), even smaller proportions can bias survey estimates and severely compromise data quality (Schräpler and Wagner 2005). Concerning partially falsified interviews, Blasius and Thiessen document suspicions for multiple large-scale surveys (e.g., Blasius and Thiessen 2013, 2015, 2021). However, such cases are difficult to verify, which complicates estimating their frequency. To deal with interviewer falsification, survey organizations often follow a dual approach: prevention and detection. Regarding prevention, strategies are mainly driven by theoretical assumptions on interviewers' motivations to falsify. For instance, DeMatteis et al. (2020) reviewed established prevention methods in the context of the fraud triangle framework developed by Cressey (1953). Accordingly, effective measures should minimize the "[p]ressure or motivation to commit the act; [p]erceived opportunity; and [r]ationalization" (DeMatteis et al. 2020, p. 18). These include informing interviewers about the consequences of falsifying interviews, informing interviewers about monitoring and verification methods, conducting background checks when hiring interviewers, and adequate payment structures (AAPOR 2003).

Not all interviewers will be deterred from falsification by these prevention measures. Therefore, survey organizations apply several techniques to detect falsifying interviewers, such as verification (or recontact) methods, which can be conducted via letter or postcard, telephone, or face to face. The scope of the recontact ranges from asking whether the interview took place to re-interviewing the respondent. However, this approach is restricted by nonresponse, respondents' failure to remember the interview, instability of responses, and increased respondent burden and survey costs (Bredl et al. 2013). Another standard approach to detect falsifying interviewers is interviewer monitoring. This method has long been limited for face-to-face interviews, but technological advances allow for more extensive monitoring procedures during the field period (see Thissen and Myers 2016 for a detailed summary).

Various statistical tools often support the aforementioned detection methods to identify suspicious interviewers, for example, outlier detection (Schwanhäuser et al. 2022) or cluster analysis (Bredl et al. 2012; De Haas and Winker 2016). These tools are usually informed by falsification indicators, which help distinguish between real and falsified interviews (Menold and Kemper 2014; Murphy et al. 2016; Schwanhäuser et al. 2022). For example, one commonly used falsification indicator is the variation of responses within same-scaled item batteries (response differentiation), which is expected to be lower for falsified interviews than for real interviews as falsifiers presumably tend to minimize their invested effort (Menold and Kemper 2014). Although helpful, statistical methods are often data driven and are sometimes based on

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

contradictory theories regarding the expected direction of some falsification indicators.

In this study, we investigate whether falsifying interviewers can be identified by their dynamic behavior over the field period. We postulate four distinct falsifier types whom we term as steady low-effort falsifiers, who use simplistic falsification strategies; steady high-effort falsifiers, who rely on complex strategies that require more effort; learning falsifiers, who adapt their behavior over the field period; and sudden falsifiers, who abruptly switch from honest interviewing to falsification during the field period. We argue that they can be distinguished from honest interviewers as their strategies generate suspicious patterns in the data.

We use data from a large-scale survey of refugees in Germany containing verified falsifications to test whether falsifiers indeed follow the postulated strategies. Using response differentiation as an approximation of falsification effort, we employ a multilevel model with interviewer effects on the intercept, the slope of the interview sequence, and the scale. To evaluate the occurrence of the falsifier types in other survey settings, we also apply the model to cross-national survey data of the general population.

2. THEORETICAL FRAMEWORK

2.1 Interviewer Falsification as Rational Behavior

Falsifiers have often been characterized as rational actors who assess their actions' expected costs and benefits to make a decision (Kennickell 2015; Kosyakova et al. 2015; Blasius and Thiessen 2021). Expected costs of falsification include sanctions such as job loss or legal consequences. The latter is rarely relevant as it is complex to provide conclusive proof of falsification, and survey organizations seek to avoid publicity on such delicate cases (Winker 2016; Blasius and Thiessen 2021). These costs only arise if the falsification is detected. Thus, if the perceived probability of detection is low, the expected costs will also be lower. Concerning the expected benefits, falsifiers can save time as it is faster to falsify an interview than to conduct a real interview. Saving time is particularly relevant for widely used piece-rate payment schemes, where interviewers receive fixed amounts for each successfully conducted interview (Kosyakova et al. 2015; Josten and Trappmann 2016). Falsifying instead of conducting a real interview may also reduce interviewer burden and thus the cognitive effort invested in each case. Real interviews require demanding tasks such as convincing the respondent to participate, administering (potentially sensitive or awkward) questions, and recording responses (West and Blom 2017). However, whether falsifying indeed reduces the effort invested in each case depends on the effort invested in the

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

falsification, as falsifiers may develop complex falsification strategies that could exceed the effort required for a real interview.

The decision on the effort level invested in each falsification also affects the perceived probability of detection. With increasing levels of effort, the perceived probability of detection decreases, and the probability of receiving sanctions is reduced. Thus, falsifiers have to weigh the risk of detection against the effort invested in each falsification and consider that the probability of detection is also influenced by the controlling procedures implemented by the survey organization (for a detailed discussion on the survey organization's incentives and potential actions concerning falsification, we refer to Winker 2016). If the falsifiers know that the controls are only superficial, they will presumably invest little effort to avoid detection.

2.2 Distinct Types of Falsifiers

As each falsifier likely weighs the potential costs and benefits of falsification differently and perceptions of detection risk may vary, we assume that falsifiers also vary with regard to their falsification behavior. Below we postulate four potential types of falsifiers and briefly discuss how their behavior could lead to suspicious patterns in the data.

We begin with *steady low-effort falsifiers* who perceive the risk of detection or the costs in case of detection to be low. Correspondingly, steady low-effort falsifiers rely on less sophisticated falsification schemes using simplistic strategies to minimize their invested effort (Murphy et al. 2016). For example, these falsifiers may produce high item nonresponse, short interview durations, or reduced response differentiation (i.e., straightlining) and could be detected by a minimum of quality control procedures.

Steady high-effort falsifiers perceive the risk of detection or the costs in case of detection as higher compared to steady low-effort falsifiers. To reduce these expected costs, they invest greater effort in falsifying data and produce no item nonresponse, realistic interview durations, or presumably inconspicuous differentiation in Likert-scaled item batteries. They might even know from previous work experience how real respondents behave and imitate these behaviors in their falsification schemes. Simplistic quality control procedures are likely to be insufficient for identifying these falsifiers. However, strictly following the same high-effort falsification strategy may reduce variation across falsified interviews. For instance, among real respondents interviewed by the same interviewer, the response differentiation within item batteries can vary from little-to-high. If falsifiers repeatedly implement the same strategy, they will likely create suspiciously low variation in response differentiation from interview-to-interview.

For steady low- and high-effort falsifiers, we assume that falsifiers behave the same way throughout the entire field period. However, falsifiers may also

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

adapt their behavior over time. Such *learning falsifiers* might behave like steady high-effort falsifiers in the beginning of the field period but adjust their estimate of the risk of detection after learning about the quality control procedures (or lack thereof) used by the survey institute. Learning falsifiers likely reduce their falsification effort to increase the benefits of falsification when they perceive the control procedures to be poor. Such falsifiers are characterized by steadily changing values used for quality control monitoring, such as response differentiation. Methods implemented for detecting steady low- or high-effort falsifiers may not work here, as higher-effort falsifications are mixed with lower-effort falsifications, depending on the learning pace. As another alternation of steady high-effort falsification behavior, learning falsifiers could also switch from fabricating parts of the interview to blatantly fabricating entire interviews.

Lastly, some falsifiers may start falsifying at some point during the field period, for instance, because of being overwhelmed by their tasks or frustrated by the lack of respondent cooperation (Crespi 1945; Gwartney 2013). For such *sudden falsifiers*, the change point is ex ante unknown and their quality control values will resemble real interviews up until the switch to falsifying and then follow either the behavior of steady low- and high-effort or even learning falsifiers. Such falsifiers are characterized by changes in data quality measures and high variation in these measures due to the switch from interviewing to falsifying.

3. DATA

First, we test whether falsifiers follow the postulated strategies using large-scale survey data containing verified falsifications. Second, we use large-scale survey data to investigate the occurrence of the posited behaviors in other publicly available survey datasets. We note that the occurrence of real falsifications in the second data set is unknown, and thus, any possible detection of suspicious interviewers does not prove the prevalence of falsifications as this requires formal investigations.

3.1 IAB–BAMF–SOEP Survey of Refugees

The data containing verified falsifications come from the first wave of the IAB–BAMF–SOEP Survey of Refugees in Germany 2016 (version SOEP.v33) (Brücker et al. 2017). After the large influx of refugees in Germany in 2015 and 2016, the panel study was launched to gather information about this population. The multi-stage cluster sample was drawn from the Central Register of Foreign Nationals (*Ausländerzentralregister*; AZR). In addition to the selected anchor person, all household members older than 18 were interviewed, if possible. The interview consisted of a household

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

questionnaire posed to the head of household (usually the anchor person) and a person questionnaire posed to every adult household member (at least 18 years old). Due to the special target population, the questionnaires were available in seven languages, both in written and audio form. Moreover, the interviewer could call a translator for assistance.

In total, 4,816 persons in 2,554 households were interviewed from June to December 2016 by 98 trained interviewers using CAPI (household level response rate 2, AAPOR 2016: 50.0 percent; Kroh et al. 2017). The interviewers received piece-rate wages for every successful interview and conducted 50 personal interviews, on average (median = 31.5, maximum workload: 289 interviews). At the beginning of the field period for the second wave, the survey organization found irregularities for respondents assigned to one interviewer (henceforth called interviewer A) in the first wave. This interviewer was found to have falsified all of their person interviews ($n = 289$), amounting to about 6 percent of the responding sample. We use these data to test whether the falsifier followed one of the four postulated strategies. Note that the affected observations were immediately removed from the data release (IAB 2017).

3.2 European Social Survey

We evaluate the extent to which the posited behavioral patterns are present in survey data that do not contain verified falsifications using data from the 6th round of the European Social Survey (ESS) (ESS Round 6: European Social Survey Round 6 Data 2012), as previous research found sizable interviewer effects on indicators of data quality for these data (Loosveldt and Beullens 2017). Furthermore, Blasius and Thiessen (2021) analyzed ESS data using methods specifically targeting partial falsifications (namely, Categorical Principal Component Analysis) and provided evidence of fraudulent interviewer behavior, though this behavior could not be conclusively verified.

The ESS is a biennial cross-sectional face-to-face survey conducted in multiple countries (for details on the survey and sampling procedures, refer to European Social Survey 2012). In 2012 and 2013, 29 countries participated in the 6th round. As an analysis of all countries participating in the ESS exceeds the scope of this paper, only data for Denmark, Hungary, and Ireland are used. These countries were selected based on Loosveldt and Beullens (2017), who found very small interviewer effects for Denmark and larger effects for Hungary and Ireland. Hence, analyzing these three countries provides a comprehensive overview on the diversity of interviewer behavior and effects in the ESS. In all three countries, the interviews were conducted via CAPI and the interviewers received piece-rate wages. Denmark obtained a response rate (AAPOR 2016, RR1) of 56.7 percent with a final sample size of 1,650 respondents, while Hungary and Ireland had response rates close to 65 percent

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

(65.1 and 65.0) (AAPOR 2016, RR1) with final sample sizes of 2,014 and 2,628, respectively (European Social Survey 2012; Beullens et al. 2014). The average number of interviews conducted per interviewer was substantially lower in the ESS samples (Denmark: 15.6; Hungary: 13.0; Ireland: 22.3) than in the IAB–BAMF–SOEP Survey of Refugees.

3.3 Dependent Variable

To approximate the falsifier's effort, we rely on a measure of response differentiation for item batteries using the same response scale (Yan 2008). We use response differentiation for two reasons. First, response differentiation is closely related to effort. Less response differentiation implies more similar responses, thereby reducing the cognitive effort of the answering process (Menold et al. 2013; Menold and Kemper 2014). Hence, less differentiation allows for faster completion of the questionnaire (and, in the case of the IAB–BAMF–SOEP Survey of Refugees and the ESS, a higher hourly interviewer wage), if the interviewer chooses the same response options regardless of their content. Second, the IAB–BAMF–SOEP Survey of Refugees questionnaire contains many long item batteries distributed over the entire questionnaire; thus, a measure based on these items will likely provide more detailed insights on the falsifiers' strategy than measures based on few questions in specific sections of the questionnaire.

Low response differentiation implies saving time and effort, thereby increasing the expected benefits of falsification. At the same time, lack of response differentiation may increase the perceived probability of detection as reduced differentiation is a suspicious response pattern, which leads to an increase in the expected costs. Therefore, when generating artificial responses to item batteries, falsifiers must take the outlined tradeoff into account which may result in distinct patterns over the field period for the proposed falsifier types.

As a robustness check and to illustrate the potential application of our approach in surveys lacking item batteries, we also use two further data quality measures: the share of rounded responses for numerical questions (Menold and Kemper 2014) and the share of extreme responses to Likert-scaled questions (Schäfer et al. 2005). Extreme responding has a looser relation to effort, and numeric questions are less frequent than item batteries in the questionnaire. Therefore, we will only briefly discuss their results and implications. Their measurement and the involved variables are described in the supplementary data S1 online.

We measure response differentiation for each interview by calculating the standard deviation of responses for several batteries of same-scaled items (following Kemper and Menold 2014). Although various approaches to measure response differentiation exist (Loosveldt and Beullens 2017; Kim et al. 2019), we use the standard deviation due to its simplicity and the possibility of

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

capturing differences on a continuous scale. A low standard deviation indicates little differentiation. As the questionnaire contains multiple appropriate item batteries, we obtain multiple standard deviations per interview. To combine the measures, we first standardize the standard deviation of every item battery across all interviews in the survey. The standardization prevents undesired effects caused by differences in scaling across item batteries. Next, the standardized standard deviations within every interview are averaged, with each standard deviation receiving a relative weight based on the number of items answered (without item nonresponse) in the respective battery and the total number of answered items across all batteries. This ensures that standard deviations calculated for longer item batteries receive a higher weight than shorter item batteries. The resulting formula is:

$$D_{ij} = \frac{\sum_{k=1}^{K} N_{ijk} \mathrm{SD}_{ijk}}{\sum_{k=1}^{K} N_{ijk}}, \tag{1}$$

where N_{ijk} is the number of answered items for item battery k in interview j by interviewer i, SD_{ijk} is the respective z-standardized standard deviation, and the denominator is the total number of answered items across all batteries. For observations with average standard deviations for all item batteries, D_{ij} is close to zero. Observations with low standard deviations have values below zero, whereas observations with high standard deviations have positive D_{ij} values. Note that we cannot establish universal thresholds that denote whether D_{ij} is too low or high, as its values depend on the number of used item batteries and their content. For example, for independent standard normally distributed random variables, the standard deviation of their sum is the square root of the number of variables. Thus, determining outlier thresholds based on variance measures depends on the number of variables. In our application, this is further complicated by correlations between variables.

The IAB–BAMF–SOEP Survey of Refugees person questionnaire includes eight appropriate item batteries with at least five items and a minimum of five response options that are used in the analysis (see table S3 in the supplementary data online for the complete list of item batteries). Batteries with fewer items or response options are not considered here to allow for finer detection of differentiation tendencies. These item batteries come from the person questionnaire (TNS Infratest Sozialforschung 2016). Due to item nonresponse, none of the item batteries was answered by all respondents. As the standard deviations are standardized, we can include observations with missing standard deviations for some item batteries in the analysis. For five respondents, the standard deviation is missing for all item batteries. These observations were

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

excluded from the analysis. The distribution of the resulting indicator is displayed in figure S3 in the supplementary data online.

For the ESS data, we use six item blocks (Loosveldt and Beullens 2017), which are listed in table S5 in the supplementary data online. Each item block contains at least five items and response options ranging from 0 to 10 or 1 to 5. The final measure of response differentiation is calculated in the same way as for the IAB–BAMF–SOEP Survey of Refugees. The distribution of the indicator is shown in figure S4 in the supplementary data online.

4. MODELING APPROACH

To test whether interviewers show suspicious behaviors over the field period, we employ multilevel modeling to disentangle interviewer from respondent effects and exploit the hierarchical data structure (respondents nested within interviewers) (Hox et al. 1991; Hox 1994). Such models have been applied to investigate interviewer effects on a variety of data quality measures (e.g., Pickery and Loosveldt 2004; Schnell and Kreuter 2005; Olson and Peytchev 2007; Kosyakova et al. 2015; Brunton-Smith et al. 2017; Loosveldt and Beullens 2017; Sharma and Elliott 2020; Sturgis et al. 2021). Among these studies, the effect of interview sequence (or within-survey experience) has also been considered (Olson and Peytchev 2007; Olson and Bilgen 2011; Kosyakova et al. 2015, 2022; Josten and Trappmann 2016; Loosveldt and Beullens 2017). However, these studies focused on overall interviewer effects and did not test whether suspicious individual interviewers can be detected. Moreover, only Brunton-Smith et al. (2017) and Sturgis et al. (2021) analyzed interviewer effects on residual variance. Pickery and Loosveldt (2004) and Sharma and Elliott (2020) are the closest to the present analysis as they use multilevel modeling to detect "exceptional" interviewers (i.e., interviewers with unusual response patterns).

We are interested in differences in the intercept, differences in the slope of the interview sequence, and differences in the residual variance across interviewers. While previous studies examined these differences in separate models, we fit a single model that contains interviewer effects on the intercept, slope, and scale. The base specification of this model is formalized below:

$$D_{ij} = \beta_0 + \theta_{i0} + (\beta_1 + \theta_{i1})\log t_{ij} + \varepsilon_{ij}, \qquad (2)$$

$$\log(\sigma_\varepsilon) = \alpha_0 + \theta_{i2}.$$

The dependent variable in the first line of the model (location equation) is response differentiation D_{ij}, which is calculated using equation (1) for each interviewer i and interview j. The interview sequence variable t_{ij} is generated

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

by sorting the interviews available for each interviewer by date and time and assigning increasing values starting at 1 to each interview for each interviewer. We use the logarithm of the interview sequence as we expect that the change in effort due to learning decreases over the field period. β_0 denotes the constant, β_1 the population parameter of the logarithm of the interview sequence, and ε_{ij} denotes the residual. θ_{i0} is the interviewer-specific effect on the intercept, and θ_{i1} is the interviewer-specific slope effect. θ_{i0}, θ_{i1}, and ε_{ij} are assumed to be independent and normally distributed with mean zero and variances $\sigma^2_{\theta_0}$, $\sigma^2_{\theta_1}$, and σ^2_{ε}, respectively. In the second line of the model (scale equation), the standard deviation of the residuals σ_ε is modeled. The standard deviation of the residuals is assumed to be log-normally distributed to ensure positive variances (Hedeker and Nordgren 2013). α_0 denotes a constant and θ_{i2} is the interviewer component of the scale equation, which is assumed to be normally distributed with mean zero and variance $\sigma^2_{\theta_2}$.

For steady low-effort falsifiers, intercept and scale effects are the key parameters. For response differentiation, steady low-effort falsifiers should have exceptionally low values, as low intercept effects indicate low response differentiation and correspondingly low effort. Low scale effects indicate that the falsifier repeatedly followed the same (low-effort) strategy. These effects should be randomly distributed around the population parameters β_0 and α_0 for honest interviewers. For steady high-effort falsifiers, only the scale effects are crucial: scale effects denote the residual variance, which is expected to be low for high-effort falsifiers who steadily follow the same strategy. For learning falsifiers, the interviewer slope effects θ_{i1} are the key parameters as they indicate interviewer-specific deviations from the population effect β_1 of the logarithmized interview sequence. For honest interviewers, θ_{i1} is expected to be close to zero, as response differentiation should not depend on the interview sequence (Kosyakova et al. 2022). For learning falsifiers, we expect a negative effect that indicates a decrease in response differentiation and thus falsification effort over the field period. Lastly, both slope and scale effects are relevant for sudden falsifiers, as changes from honest interviewing to falsification should result in a change in response differentiation. Whether the deviation is positive or negative depends on the sudden falsifier's strategy. For the scale effects, positive deviations should flag sudden falsifiers as the switch to falsifications should result in increased residual variance.

Across the three types of interviewer effects, we apply the same rules for deeming interviewers suspicious. First, their credible interval for the respective effect must not include zero. Second, they must have posterior means exceeding the boxplot whiskers (25th/75th percentile ± 1.5 times the interquartile range) for the distribution of the posterior medians. Note, however, that the defined outlier rule based on boxplot whiskers may lead to false positives, and alternative outlier rules may lead to different results. Therefore, flagged interviewers should be investigated case-by-case.

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

We note that interviewers employed in the IAB–BAMF–SOEP Survey of Refugees and the ESS were not randomly assigned to households across the respective countries but were assigned to regional clusters: primary sampling units. Therefore, the results observed for interviewers may be driven by regional clustering effects (Schnell and Kreuter 2005). To disentangle interviewer and regional cluster effects, we would require sufficient interpenetration, that is, interviewers must work in multiple clusters, and multiple interviewers must work in a given cluster, which is not prevalent in the data used here. For the IAB–BAMF–SOEP Survey of Refugees, however, regional clusters are expected to have only small effects, as the target population are recently arrived refugees subject to state-based residential allocation policies following a political quota (BAMF 2019; Kosyakova et al. 2019). Lastly, controlling for small-scale regional effects could prevent the detection of falsifiers operating or cooperating in the same region (Yamamoto and Lennon 2018; Bergmann et al. 2019).

Nonetheless, we test the robustness of the results by including control variables for respondent and area characteristics in the model for the IAB–BAMF–SOEP Survey of Refugees and ESS data. They include respondents' age, gender, education, living arrangement (only for the IAB–BAMF–SOEP Survey of Refugees), an urban-rural binary variable, as well as federal state/region fixed effects (see supplementary data S4 online). Note that the included variables are more likely to explain overall effects on the dependent variable than extreme slopes or scales observed for single interviewers.

The model is fitted using Markov chain Monte Carlo (MCMC) methods. In particular, we use the No-U-Turn Sampler (Hoffman and Gelman 2014), a version of the Hamiltonian Monte Carlo algorithm implemented in Stan (Carpenter et al. 2017) and accessed via the brms interface (Bürkner 2017, 2018) in R (R Core Team 2020). The model is fitted using eight chains of 8,000 iterations, each with a burn-in period of 3,000 iterations. We specify flat priors for the population-level coefficients and default half student-t priors with three degrees of freedom for the standard deviations of the interviewer effects. We assessed whether the priors for the interviewer effects affect the results by trying different priors such as half Cauchy and inverse Gamma distributions, but the results did not change. Model convergence was evaluated by the $\widehat{R}$ statistic with a critical value of 1.01 for each model parameter (Gelman et al. 2013, p. 285) and by ensuring that there were no divergent transitions (Betancourt 2017). Estimates and credible intervals shown in the results section are based on posterior distributions for the model parameters obtained from the MCMC draws.

5. RESULTS

For each dataset, we first estimate the base specifications with no control variables and then conduct a robustness check with control variables. Note that we

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

are not interested in explaining differences in levels, slopes, or residual variances but in detecting suspicious interviewers.

5.1 Analysis of the IAB–BAMF–SOEP Survey of Refugees

The interviewer effects for the IAB–BAMF–SOEP Survey of Refugees are displayed in figure 1. Figure 1a shows the effects on the intercept, figure 1b shows the effects on the slope, and figure 1c shows the effects on the scale. In each panel, each point corresponds to a single interviewer. The interviewer effects are sorted by size, and 95 percent credible intervals are provided. The dashed horizontal lines depict the boxplot whiskers. The estimation results are reported in column 1 in table S9 in the supplementary data online. The estimated coefficient of the logarithmized interview sequence is positive but negligible in size. From the first to the 10th interview, response differentiation increases by 0.058, which equals roughly 12 percent of one standard deviation. Hence, overall the response differentiation changes only slightly over the field period.

Figure 1a shows that the verified falsifier's intercept effect is not suspicious (ranked 68th). The first-ranked interviewer (interviewer D) is suspicious but conducted only eight interviews. The last-ranked interviewer (interviewer E) has rather high differentiation values and conducted 27 interviews. None of these interviewers was flagged by further statistical identification methods (see Kosyakova et al. 2019) or further checks (such as recontacts) by the survey organization. Thus, relying on the intercept effects alone is insufficient for identifying the verified case. Remember that the intercept effects denote differences at the first interview as interview sequence effects are included in the model.

As displayed in figure 1b, most of the slope effects for the interview sequence are close to zero, or the credible intervals include zero. As with the intercept effects, interviewer A does not deviate from the other interviewers and is ranked 28th. For the first-ranked interviewer (henceforth called interviewer B), however, the slope value deviates substantially from the others, implying a decrease in response differentiation over the field period. Similarly, the second-ranked interviewer (henceforth called interviewer C) is suspicious with a negative slope effect. Interviewers B and C conducted 46 and 16 interviews, respectively. Accordingly, interviewers B and C reveal a suspicious slope effect consistent with learning behavior or switching from honest interviewing to falsification. These results and conclusions drawn from further statistical checks were reported to the survey organization, who verified that interviewers B and C were indeed deviant, although the survey organization could not exactly tell which interviews were falsified (Kosyakova et al. 2019). The published data were immediately revised after the detection (IAB et al. 2019).

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

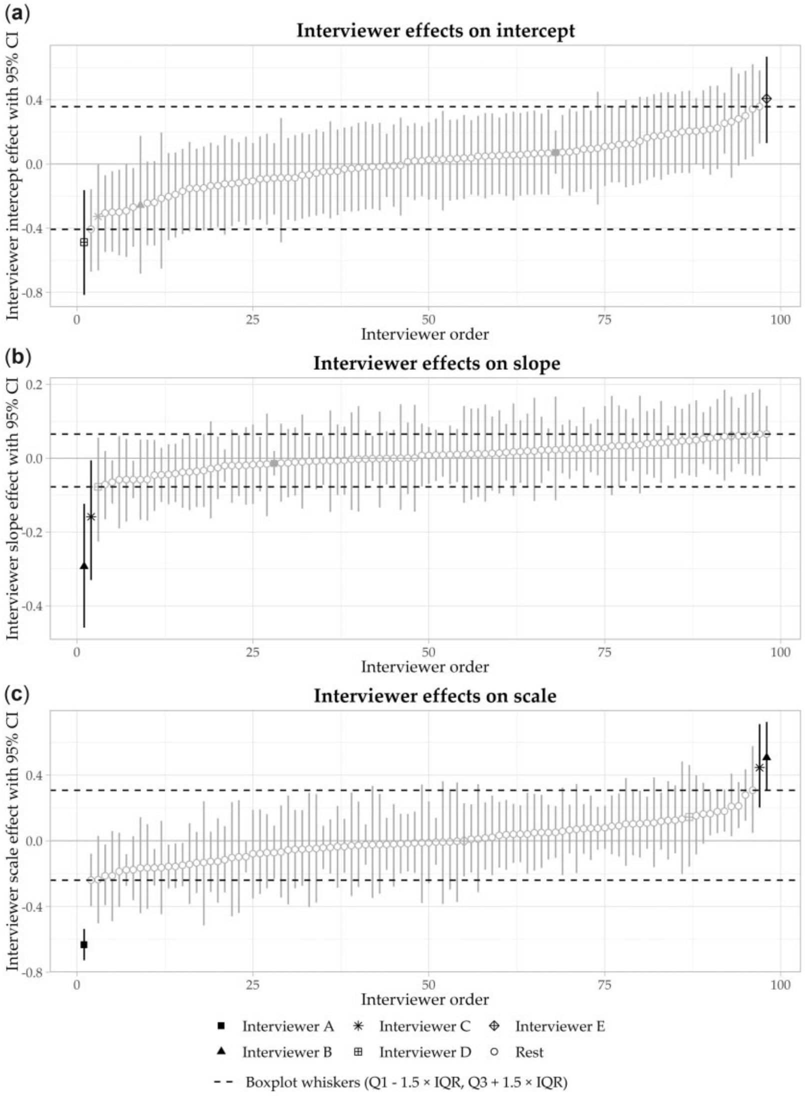

Figure 1. Interviewer Effects on Intercept, Slope, and Scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on model 1 in table S9 in the supplementary data online. IAB-BAMF-SOEP Survey of Refugees 2016.

Lastly, figure 1c shows that the scale effects are distributed homogeneously, except for the first-ranked interviewer and the last-ranked pair of interviewers. The first-ranked interviewer is interviewer A who has a suspiciously low scale effect. This indicates limited variation in response differentiation,

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

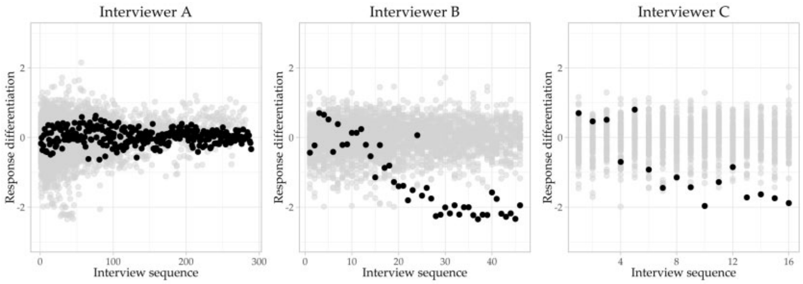

Figure 2. Development of Response Differentiation for Verified Falsifiers. Black dots correspond to the respective falsifier, and gray dots correspond to the rest of the sample. IAB-BAMF-SOEP Survey of Refugees 2016.

which—combined with the inconspicuous intercept effect—is expected for steady high-effort falsifiers. The two last-ranked interviewers are interviewers B and C. The exceptional slope effects observed for these interviewers also lead to a suspicious scale effect.

To investigate the behaviors of interviewer A and the additionally identified interviewers B and C in greater detail, figure 2 shows the development of response differentiation over the field period for each of them. For reference, the differentiation values for the rest of the sample are also shown. Interviewer A has a relatively low variation in response differentiation with values around zero, close to the overall average in the sample. Such a pattern is in line with the behavior expected of a steady high-effort falsifier. In contrast, interviewer B has relatively high response differentiation values at the beginning of the field period that steadily decrease from the 10th interview onward. Toward the end of the field period, response differentiation is clearly below the "normal" values observed for the rest of the sample. This pattern is in line with a learning falsifier, although it is also possible that some of the first interviews were real, and the interviewer switched to falsification. For interviewer C, the pattern is less clear due to the limited number of available observations. The response differentiation values are at the upper end of the distribution in the beginning of the field period, but this strictly changes after the 5th interview. This break may either illustrate learning behavior or a change from conducting real interviews to falsifying interviews. As mentioned above, detailed information on whether every interview of interviewers B and C was falsified is not available.

To test the robustness of the results, we replicate the benchmark models by adding multiple control variables (gender, age, education, accommodation, region, rural/urban) to the location equation. The estimation results of these models are reported in table S10 in the supplementary data online. As illustrated in figure S5 in the supplementary data online, the deviations of interviewers B and C for the slope effects cannot be explained by the included explanatory variables, although the effect for interviewer C is now closer to

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

zero. Similarly, the explanatory variables cannot explain the deviation of interviewer A for the scale effects.

Figures S9 and S10 in the supplementary data online show the results for the model without controls for two further indicators, extreme responding and rounding. For extreme responding, two interviewers are flagged for the intercept effects, but their values are close to the rest of the sample. Interviewers A, B, and C have inconspicuous values. Interviewers B and C have exceptionally negative slope effects, although interviewer C is barely below the boxplot rule. For the scale effects, one interviewer is flagged, and interviewer A is ranked second, but their values seem in line with the distribution for the remaining sample. For rounding, interviewer A has a suspiciously low intercept effect, and interviewer B has a suspiciously large intercept effect. The slope effects depict that several interviewers have negative effects, but these values are not particularly exceptional. Interviewer B is the only interviewer with a relatively large slope effect, denoting that the share of rounded responses increased over the field period. A closer inspection revealed that this interviewer heavily reduced the number of valid responses to open-ended numeric questions over the field period, which led to frequent high shares of rounded answers as, for example, one numeric item in the interview had a valid response, and this response was a rounded number, which results in a share of 100 percent. For the scales, interviewer A has an exceptionally negative value, indicating reduced residual variance. Interestingly, interviewer C is not flagged by any of the estimated interviewer effects. In summary, for neither of the indicators are all three interviewers A, B, and C flagged.

5.2 Analysis of the European Social Survey

For the ESS data, we only discuss the results for Ireland in detail while touching on the results for Denmark and Hungary only briefly for brevity. Figure 3 displays the interviewer effects for Ireland. We also fit the multilevel model for the three countries with covariates, and the results remain robust to these extensions (see supplementary data S5 and S6 online). Figure 3a shows that multiple interviewers have suspiciously low or high values of response differentiation, although most of them do not significantly differ from the unflagged interviewers. As displayed in figure 3b, there is an interviewer who has a suspicious negative slope effect indicating potential learning behavior. One further interviewer has a suspicious positive slope effect. Finally, figure 3c shows two interviewers with suspicious negative scale effects expected of steady high- or low-effort falsifiers. The first-ranked interviewer is the interviewer who is ranked first in figure 3a, which is expected for a steady low-effort falsifier. None of the other interviewers has multiple suspicious effects.

Next, we take a closer look at the development of response differentiation over the field period for the flagged interviewers. Figure 4 shows the

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

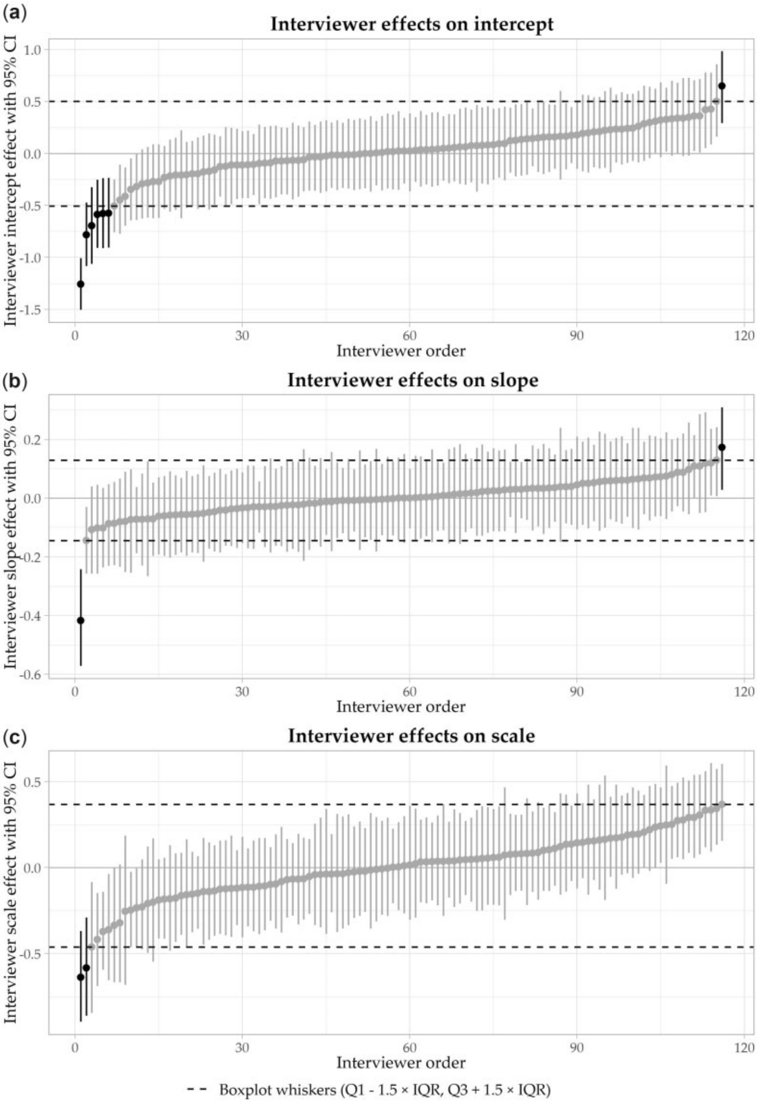

Figure 3. Interviewer Effects on Intercept, Slope, and Scale. Bolded interviewers have median posteriors exceeding the boxplot whiskers and credible intervals not including zero. Predictions are based on model 3 in table S9 in the supplementary data online. ESS Round 6, Ireland.

differentiation results for the interviewer ranked first for the slope effects (interviewer ESS-A), and the interviewers ranked first and second for the scale effects (interviewers ESS-B and ESS-C, respectively). For interviewer ESS-A, response differentiation decreases over the field period and thus follows the

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

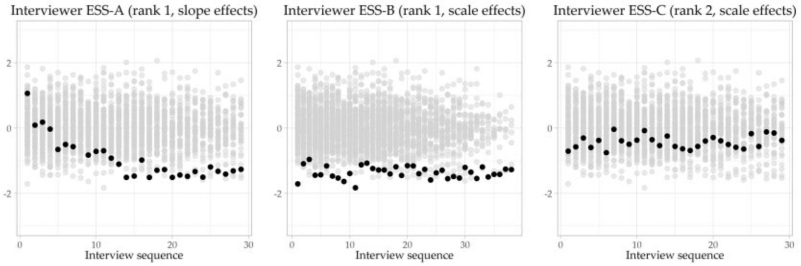

Figure 4. Development of Response Differentiation. Black dots correspond to the respective suspicious interviewer, and gray dots correspond to the rest of the sample. ESS Round 6, Ireland.

behavior expected of a learning falsifier. Interviewer ESS-B—ranked first for the intercept and scale effects—shows limited variation around reduced response differentiation, which is characteristic of a steady low-effort falsifier. Lastly, interviewer ESS-C also has limited variation around close to average response differentiation, suggesting deviant behavior consistent with a more sophisticated falsifier.

The results for Denmark and Hungary are provided in supplementary data S8 online. For Denmark, we find only minor interviewer effects on the intercept, slope, and scale and no suspicious interviewer, which is in line with previous research on interviewer effects in Scandinavian countries (Loosveldt and Beullens 2017). We find more evidence of interviewer effects in the Hungary sample. Multiple interviewers have negative effects on the intercept, although none of these effects is suspicious. There are several interviewers with slope effects below the boxplot whisker line, but only for one interviewer who does not significantly differ from unsuspicious interviewers does the credible interval not include zero. With regard to the scale effects, several interviewers have rather low values indicative of behavior expected of steady high- or low-effort falsifiers, but none of the effects exceeds the boxplot whisker rule. Although the interviewer effects are not as suspicious as for the IAB–BAMF–SOEP Survey of Refugees or Ireland, interviewers with particularly low intercept, slope, or scale effects may have required closer inspection.

6. DISCUSSION

Falsified interviews can substantially bias survey results (e.g., Schräpler and Wagner 2005). To prevent and detect falsifications, survey methodologists should not only use empirical detection methods, but also comprehend falsifiers' motivations and behaviors. In this study, we posited four distinct falsifier types: steady low-effort falsifiers, steady high-effort falsifiers, learning falsifiers, and sudden falsifiers. Using data containing verified falsifications and

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

multilevel models, we retrospectively identified a presumably steady high-effort falsifier previously detected by the survey organization based on their behavior over the field period. In addition, the method identified two further interviewers with suspicious behavior expected of learning and sudden falsifiers, who were later confirmed as deviant by further statistical analyses and recontact checks performed by the survey organization. Altogether, these results emphasize the importance of taking a variety of potential motivations and falsification strategies into account when analyzing deviant interviewer behavior. Our analysis of the ESS data shows that such behavior also appears in other publicly available datasets. Note, however, that only formal investigations can prove falsifications.

Survey practitioners may add the presented methods to their general data quality control procedures. First, graphical tools similar to figures 2 and 4 can be applied to monitor interviewers during the field period. Second, applying the multilevel model to survey data after the field period or when interviewers have conducted a reasonable number of interviews can provide useful insights into interviewers' behavior. Of course, applying the model when the number of interviews per interviewer is still small will provide limited insights. For example, it is difficult to identify outlying slope or scale effects for interviewers with only five interviews. Instead, practitioners may start with simpler versions of the model, such as intercept-only multilevel models that allow for identifying the most blatant falsifiers early in the field period. With sufficient data per interviewer, the more complex model can be used to identify more sophisticated falsifiers. In such applications, the models may identify both partial and complete fabricators, although we could not evaluate the method's effectiveness for partial fabrications due to a lack of verified data. In any case, falsifiers should be detected as early as possible to facilitate formal investigations.

Nevertheless, four caveats remain. First, the method's efficiency depends on the selected outcome variable. Hence, researchers must carefully select appropriate data quality indicators depending on the questionnaire content. Second, some of the postulated falsifier types are easier to detect than others. For example, low-effort falsifiers will always leave obvious traces in the data. To the contrary, in some cases, very sophisticated falsifiers may even outsmart the complex multilevel modeling approach, although it seems unlikely that a falsifier knows both the mean and the variance of data quality measures ex ante. Third, detecting trends for interviewers with a limited number of observations (e.g., <10) is challenging, which is relevant for learning and sudden falsifiers. Lastly, suspicious interviews are identified on the interviewer level, which is why single falsified interviews cannot be identified using this method. Future research may address these limitations, for example, by using data with verified falsified interviews and detailed paradata allowing for more fine-grained analyses. However, the release of publicly available data containing verified falsified interviews is rare. Thus, we encourage survey organizations to make

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

such data available to researchers to help advance our understanding of interviewer falsification.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

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

AAPOR (2016), "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys," 9th edition, AAPOR.

BAMF (2019), "Initial Distribution of Asylum-Seekers." http://www.bamf.de/EN/Fluechtlingsschutz/AblaufAsylv/Erstverteilung/erstverteilung-node.html.

Bergmann, M., Schuller, K., and Malter, F. (2019), "Preventing Interview Falsifications during Fieldwork in the Survey of Health, Ageing and Retirement in Europe (SHARE)," *Longitudinal and Life Course Studies*, 10(4), 513–530.

Betancourt, M. (2017), "A Conceptual Introduction to Hamiltonian Monte Carlo." http://arxiv.org/abs/1701.02434

Beullens, K., Matsuo, H., Loosveldt, G., and Vandenplas, C. (2014), *Quality Report for the European Social Survey, Round 6*. London: European Social Survey ERIC.

Blasius, J., and Thiessen, V. (2013), "Detecting Poorly Conducted Interviews," in *Interviewers' Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 67–88.

———. (2015), "Should We Trust Survey Data? Assessing Response Simplification and Data Fabrication," *Social Science Research*, 52, 479–493.

———. (2021), "Perceived Corruption, Trust, and Interviewer Behavior in 26 European Countries," *Sociological Methods and Research*, 50(2), 740–777.

Bredl, S., Storfinger, N., and Menold, N. (2013), "A Literature Review of Methods to Detect Fabricated Survey Data," in *Interviewers' Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 3–24.

Bredl, S., Winker, P., and Kötschau, K. (2012), "A Statistical Approach to Detect Interviewer Falsification of Survey Data," *Survey Methodology*, 38(1), 1–10.

Brücker, H., Rother, N., and Schupp, J. (2017), "IAB-BAMF-SOEP Befragung von Geflüchteten 2016: Studiendesign, Feldergebnisse Sowie Analysen zu Schulischer wie Beruflicher Qualifikation, Sprachkenntnissen Sowie Kognitiven Potenzialen," *IAB-Forschungsbericht*, 13, 1–76.

Brunton-Smith, I., Sturgis, P., and Leckie, G. (2017), "Detecting and Understanding Interviewer Effects on Survey Data by Using a Cross-Classified Mixed Effects Location–Scale Model," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2), 551–568.

Bürkner, P. C. (2017), "brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, 80(1), 1–28.

———. (2018), "Advanced Bayesian Multilevel Modeling with the R Package brms," *R Journal*, 10(1), 395–411.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76(1), 1–32.

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

Cressey, D. R. (1953), *Other People's Money*, Montclair, NJ: Patterson Smith.

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

De Haas, S., and 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.

DeMatteis, J. M., Young, L. J., Dahlhamer, J., Langley, R. E., Murphy, J., Olson, K., and Sharma S. (2020), *Falsification in Surveys*, Washington, DC: American Association for Public Opinion Research.

ESS Round 6: European Social Survey Round 6 Data (2012), “Data File Edition 2.4. NSD—Norwegian Centre for Research Data, Norway—Data Archive and Distributor of ESS Data for ESS ERIC.” https://doi.org/10.21338/NSD-ESS6-2012

European Social Survey (2012), “ESS6—2012 Documentation Report: The ESS Data Archive, Edition 21,” pp. 1–221.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis* (3rd ed.), New York: CRC Press.

Gwartney, P. A. (2013), “Mischief versus Mistakes: Motivating Interviewers to Not Deviate,” in *Interviewers’ Deviations in Surveys—Impact, Reasons, Detection and Prevention*, eds. P. Winker, N. Menold, and R. Porst, Frankfurt am Main: Peter Lang, Academic Research, pp. 195–215.

Hedeker, D., and Nordgren, R. (2013), “MIXREGLS: A Program for Mixed-Effects Location Scale Analysis,” *Journal of Statistical Software*, 52(12), 1–38.

Hoffman, M. D., and Gelman, A. (2014), “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,” *Journal of Machine Learning Research*, 15, 1593–1623.

Hox, J. J. (1994), “Hierarchical Regression Models for Interviewer and Respondent Effects,” *Sociological Methods & Research*, 22(3), 300–318.

Hox, J. J., de Leeuw, E. D., and Kreft, I. G. G. (1991), “The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model,” in *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, New Jersey: John Wiley & Sons, Inc, pp. 439–461.

IAB (2017), “Revidierter Datensatz der IAB-BAMF-SOEP Befragung von Geflüchteten.” doku.iab.de/grauepap/2017/Revidierter_Datensatz_der_IAB-BAMF-SOEP-Befragung.pdf.

IAB, BAMF, and SOEP (2019), “Qualitätsprüfung der Daten der IAB-BAMF-SOEP Befragung von Geflüchteten.” http://doku.iab.de/fdz/iab_bamf_soep/IAB-BAMF-SOEP_Statement_Qualitaetskontrollen_DE.pdf.

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

Kemper, C. J., and Menold, N. (2014), “Nuisance or Remedy? The Utility of Stylistic Responding as an Indicator of Data Fabrication in Surveys,” *Methodology*, 10(3), 92–99.

Kennickell, A. B. (2015), “Curbstoning and Culture,” *Statistical Journal of the IAOS*, 31(2), 237–240.

Kim, Y., Dykema, J., Stevenson, J., Black, P., and Moberg, D. P. (2019), “Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail—Web Mixed-Mode Surveys,” *Social Science Computer Review*, 37(2), 214–233.

Kosyakova, Y., Olbrich, L., Sakshaug, J. W., and Schwanhäuser, S. (2019), “Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany.” FDZ-Methodenbericht 2.

———. (2022), “Positive Learning or Deviant Interviewing? Mechanisms of Experience on Interviewer Behavior,” *Journal of Survey Statistics and Methodology*, 10(2), 249–275.

Kosyakova, Y., Skopek, J., and Eckman, S. (2015), “Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach,” *International Journal of Public Opinion Research*, 27(3), 417–431.

Kroh, M., Kühne, S., Jacobsen, J., Siegert, M., and Siegers, R. (2017), “Sampling, Nonresponse, and Integrated Weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4).” SOEP Survey Papers 477. Berlin: DIW

Loosveldt, G., and Beullens, K. (2017), “Interviewer Effects on Non-Differentiation and Straightlining in the European Social Survey,” *Journal of Official Statistics*, 33(2), 409–426.

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024

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

Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O., and Hsieh, Y. P. (2016), "Interviewer Falsification: Current and Best Practices for Prevention, Detection, and Mitigation," *Statistical Journal of the IAOS*, 32(3), 313–326.

Olson, K., and Bilgen, I. (2011), "The Role of Interviewer Experience on Acquiescence," *Public Opinion Quarterly*, 75(1), 99–114.

Olson, K., and Peytchev, A. (2007), "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes," *Public Opinion Quarterly*, 71(2), 273–286.

Pickery, J., and Loosveldt, G. (2004), "A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers," *Journal of Official Statistics*, 20(1), 77–89.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.r-project.org/.

Robbins, M. (2019), "New Frontiers in Detecting Data Fabrication," in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, eds. T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, and B. Dorer, John Wiley & Sons, pp. 771–805.

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

Schnell, R., and Kreuter, F. (2005), "Separating Interviewer and Sampling-Point Effects," *Journal of Official Statistics*, 21(3), 389–410.

Schräpler, J.-P., and Wagner, G. 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(1), 7–20.

Schwanhäuser, S., Sakshaug, J. W., and Kosyakova, Y. (2022), "How to Catch a Falsifier: Comparison of Statistical Detection Methods for Interviewer Falsification," *Public Opinion Quarterly*, 81(1), 1–31.

Sharma, S., and Elliott, M. R. (2020), "Detecting Falsification in a Television Audience Measurement Panel Survey," *International Journal of Market Research*, 62(4), 432–448.

Sturgis, P., Maslovskaya, O., Durrant, G., and Brunton-Smith, I. (2021), "The Interviewer Contribution to Variability in Response Times in Face-to-Face Interview Surveys," *Journal of Survey Statistics and Methodology*, 9(4), 701–721.

Thissen, M. R., and Myers, S. K. (2016), "Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality," *Statistical Journal of the IAOS*, 32(3), 339–347.

TNS Infratest Sozialforschung (2016), "Erhebungsinstrumente der IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Integrierter Personen- und Biografiefragebogen, Stichproben M3-M4." SOEP Survey Papers (Series A): 362, Berlin: DIW/SOEP.

Weisberg, H. F. (2005), *The Total Survey Error Approach: A Guide to the New Science of Survey Research*, Chicago: The University of Chicago Press.

West, B. T., and Blom, A. G. (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5(2), 175–211.

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.

Yamamoto, K., and Lennon, M. L. (2018), "Understanding and Detecting Data Fabrication in Large-Scale Assessments," *Quality Assurance in Education*, 26(2), 196–212.

Yan, T. (2008), "Nondifferentiation," in *Encyclopedia of Survey Research Methods*, ed. P. J. Lavrakas, Thousand Oaks, CA: SAGE Publications, Inc, pp. 520–521.

Downloaded from https://academic.oup.com/jssam/article/12/1/14/6967120 by UB Bamberg user on 03 May 2024