FALSIFICATION IN SURVEYS

Prepared for AAPOR Council and the Executive Committee of the American Statistical Association by the members of the Task Force on Data Falsification:

Jill M. DeMatteis, *Westat, Task Force Co-Chair*
Linda J. Young, *National Agricultural Statistics Service, Task Force Co-Chair*
James Dahlhamer, *National Center for Health Statistics*
Ronald E. Langley, *University of Kentucky*
Joe Murphy, *RTI International*
Kristen Olson, *University of Nebraska-Lincoln*
Sharan Sharma, *University of Maryland*

This report was commissioned by the AAPOR Executive Council as a service to the profession. The report was reviewed and accepted by AAPOR Executive Council. The opinions expressed in this report are those of the author(s) and do not necessarily reflect the views of the AAPOR Executive Council. The authors, who retain the copyright to this report, grant AAPOR a non-exclusive perpetual license to the version on the AAPOR website and the right to link to any published versions.

The task force gratefully acknowledges the contributions of Mario Callegaro, Michael Larsen, and Yan Li, who contributed to early discussions and development of the framework for this report. We also thank Rosalynn Yang, who conducted a comprehensive literature review. We also acknowledge helpful comments from AAPOR Executive Council members, ASA Executive Committee members, and their designees. The task force also gratefully acknowledges the editorial support provided by Westat.

September 2020

Table of Contents

Chaper Page

Appendixes

Tables **Page**

Figures

1. Introduction

Data fabrication and falsification pose serious threats to the credibility of survey research. Falsified or fabricated data yield biased estimates, affect the precision of estimates, and impact multivariate relationships. The Office of Research Integrity (ORI) (ORI, 2002, p. 2) declared falsification as a form of scientific misconduct. In an era of tight budgets and increasing challenges for surveys, resources still need to be allocated to prevent, detect, and mitigate data falsification and fabrication.

In considering the aspects covered by this report, it is also important to be clear about aspects not covered by the report. While data falsification and fabrication are also concerns in other fields, such as clinical trials, the focus of this report is on data falsification and fabrication in *surveys*. The types of data falsification, methods for preventing or detecting it, and the impact on analyses will be discussed for all facets of a survey, from start to finish. This report does not cover "task simplification" methods respondents use, such as stylistic response or satisficing (Blasius and Thiessen, 2015); nor does it cover outright respondent fraud, such as a single individual completing a web survey multiple times in order to collect incentives. While aspects of detecting these methods and their impacts on analyses are similar to those for falsification or fabrication by employees of the survey organization conducting the survey, the respondent's impacts on survey data quality are not explicitly covered in this report nor does it cover unintentional errors that affect data quality, such as errors in data entry or coding.

Several publicized cases of fabrication and/or falsification have drawn into question the findings of the affected surveys. As early as the 1940s, Crespi (1945) and Bennett (1948) described the "cheater problem" in interviewing. Kindred and Scott (1993) discuss falsification in the 1990 decennial census nonresponse follow-up effort. Harrison and Krauss (2002) contend that interviewer fabrication is more pervasive than generally believed. Spagat (2010) presented claims of falsification in challenging estimates of the total number of fatalities to Iraqi civilians in the Iraq war published in *The Lancet* by Burnham, Lafta, Doocy, and Roberts (2006). Broockman, Kalla, and Aronow (2015) reported their findings of irregularities in research on the effect of a brief interaction with a gay canvasser on feelings about gay people reported in *Science* by LaCour and Green (2014), resulting in a subsequent retraction at the request of Green. Significant near-duplication of records from surveys administered in Kuwait and Yemen for the Arab Barometer were reported by Robbins (2015).

Blasius and Thiessen (2015) discuss potential evidence of fabrication in the 2009 Program for International Student Assessment (PISA) study. Cases such as these have intensified concerns among survey research organizations about falsification and fabrication being a growing threat to the credibility of surveys.

While much of the focus has been on "curbstoning," or fabrication of entire interviews by interviewers, falsification goes well beyond interviewer curbstoning (U.S. Department of Health and Human Services, 2005). Falsification may only involve part of the interview or interview process, making it more challenging to detect than curbstoning given the combination of falsified and "real" data. Falsification may be done by other members of survey organizations, including supervisors, data processors, or researchers, and encompasses other falsification methods, such as altering of disposition codes, duplicating entire survey data records, or duplicating records with changes to a few answers.

Survey research organizations have developed various methods for preventing and detecting falsification and fabrication. These range from supervisor and interviewer training on the importance of integrity in the data collection effort to the use of technologies, such as computer-assisted-recorded interviewing (CARI) and global positioning system (GPS) tracking of interviewers to algorithms that analyze patterns in the data to detect duplicates, near-duplicates, and other evidence of fabrication. Actions taken by survey organizations to address falsification once detected, including reporting known falsification, are also important.

Organizations have, in many cases, developed their own protocols and procedures for preventing, detecting, investigating, and reporting on falsification; however, best practices for the field were only established in the last two decades (American Association for Public Opinion Research (AAPOR 2003; AAPOR 2005). Given the increased number of falsification and fabrication accusations, it is important that survey researchers have a good understanding of best practices surrounding the prevention and detection of falsification in surveys. To this end, as the leading professional organizations for those conducting survey research, AAPOR and the American Statistical Association (ASA) jointly engaged a task force to make recommendations on best practices for the methods to prevent and detect falsification in surveys. This most recent effort, which resulted in this report, is the culmination of efforts by these organizations to understand and address the problem of falsification in surveys. An April 2003 report "Interviewer Falsification in Survey Research:

Current Best Methods for Prevention, Detection and Repair of Its Effects," developed by representatives of survey research organizations at a summit on the topic of survey data falsification, was adopted by AAPOR and the Survey Research Methods Section (SRMS) of the ASA. A second falsification summit in 2005 resulted in the development of standards for reporting of falsification incidents, both internally and to ORI (AAPOR, 2005). AAPOR's Transparency Initiative is one effort already underway that, among other aims, helps to counter falsification.

We recognize that ASA and AAPOR are not the first organizations to address falsification and integrity in research. As part of the process of developing this report, the task force reviewed standards, guidelines, policies, and practices of other organizations and institutions. In considering the effects of claims of falsification on survey organizations themselves, we recommend that we, as a profession, are mindful about making claims of falsification writ large. As noted by Goldenring (2010), "The present [scientific integrity adjudication] system takes a position of guilty until proven innocent, a concept that is antithetical to American principles of jurisprudence. Yet this stance is acceptable as a requirement for membership in the scientific community…." Simonsohn (2013) suggests a set of measures that could be used in falsification investigations to avoid "witch-hunting."

This report begins with a discussion of the types of survey data falsification (Chapter 2). This is followed by a review of approaches for preventing falsification (Chapter 3); tools and methods for detecting falsification (Chapter 4); an examination of the impacts of falsification on study results (Chapter 5); a sampling of existing organization guidelines, policies, and procedures regarding falsification (Chapter 6); and a summary of the report (Chapter 7).

2. Types of Fabrication and Falsification

2.1 Introduction

The terms data fabrication and data falsification are closely related but are often presented as having distinctly different meanings in the literature. Definitions of data fabrication include the "invention of data or cases" (Fanelli, 2009:1); "the creation of false data by field workers" (Spagat, 2010, p. 2); and "making up data or results and recording or reporting them" (Koczela et al., 2015, p. 414). That is, data fabrication occurs when data are created and reported for some portion of an existing record or for an entire record.

Data falsification has been variously defined in the literature as "willful distortion of data or results" (Fanelli, 2009, p. 1). It is also "the creation of false data by one or more authors of the study. Falsification includes misrepresentation and suppression of other evidence relevant to the claims of that study" (Spagat, 2010, p. 2), and "manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record" (Koczela et al., 2015, p. 414).

As these definitions suggest, the distinctions between data fabrication and data falsification are often blurred. For a particular study, distinguishing between the two is challenging, if even possible. Regardless, both can lead to unreported data being used in analyses, potentially damaging the study results and conclusions. For consistency, the term falsification will be used throughout this report to encompass the full range of activities that would be considered either fabrication or falsification.

One common thread through the definitions described above is the "intentional" departure from research protocols. Intentional means the actor, whether it be an interviewer, a supervisor, or a principal investigator, is aware that his/her actions deviate from guidelines and instructions (AAPOR, 2004). Hence, falsification differs from researcher incompetence or sloppiness.

It is also clear that proving intent imposes a high burden upon survey researchers and methodologists. As an example, in reinterview (a process by which a sample of interviewer's workload is recontacted to verify their work) a discrepancy with the household roster (e.g., an adult was left off) may be identified in a single interview. Did the interviewer intentionally leave this adult off the roster? Or did the interviewer simply fail to probe for additional household members, possibly due to ineffective training? Or maybe the respondent simply forgot to report this person? If the household roster was recorded, say through the use of computer-assisted recording of interviews (CARI), answers to these questions may be more forthcoming. If CARI was not in place, the answers are more elusive. Either way, proving the interviewer intentionally left the person off the roster demands additional investigation and the use of multiple methods of falsification detection. A discrepancy of this sort should prompt review of additional cases within an interviewer's current and subsequent workloads. Discrepancies in household rosters across multiple interviews would be more suggestive of falsification. Review of the survey data may reveal a pattern in that the interviewer consistently leaves adults off the roster so that the roster respondent can be selected for a subsequent interview. As discussed in later chapters, a robust detection program relies on multiple methods to identify instances of survey falsification.

2.2 Types of Falsification Based on Steps of the Survey Process

In this section, we describe the major types of survey falsification by stage of the survey process, citing case studies that illuminate the issues. At each stage, the potential for falsification poses a threat to data integrity and quality. Depending on the survey design, the various stages may be vulnerable to the threat of falsification at the interviewer or supervisor level. Each "actor" serves a different role and has different incentives to perform with a high level of integrity or to falsify. Because the incentives vary by survey stage and actor, falsification itself may be conducted at different levels or in different ways. For instance, an interviewer may falsify at the respondent level by fabricating an entire interview (also known as "curbstoning") or may code a case as ineligible when it should have been pursued for an interview. An interviewer may also falsify data at the item level, recording erroneous data during a particularly burdensome portion of the interview or to reduce the total administration time of the interview (Murphy et al., 2016).

AAPOR (2003) reviewed factors leading to interviewers being more "at risk" for falsification. A study's design may be a contributing factor. Developing complicated and/or long questionnaires, and taking auxiliary measurements are part of an effort to maximize the information obtained from a survey. However, they not only require high response rates from reluctant populations but also can increase stress on the interviewers and their supervisors, thereby increasing the risk of falsification. If interviewers are asked to conduct surveys in regions where they fear for their personal safety, falsification may occur. In addition, organizational factors may affect the potential for interviewers to falsify data. Such factors include excessive workload, inadequate compensation, piece-rate compensation as the primary pay structure, inadequate supervision, poor quality control, off-site isolation of interviewers from the parent organization, lack of concern about interviewer motivation, and hiring and training practices that ignore falsification threats.

The remainder of this chapter presents examples of types of falsification that can and has occurred at each stage of the survey process, noting the principal actors (interviewer, field supervisor, other field staff, researcher) involved.

2.2.1 Counting and Listing (Interviewer, Field Supervisor)

Falsification may occur at the counting and listing stage. Here, interviewers, listers, or field supervisors may deliberately count an incorrect number of housing units in an area or list an inaccurate set of housing units. We note that this is not the normal human error involved in creating a list of housing units (see, for example, Eckman and Kreuter, 2013, on typical measurement and coverage errors involved in listing housing units) but a deliberate attempt to undercount or overcount housing units or to mis-assign housing units to particular streets or other geographic areas. Interviewers or supervisors may deliberately misclassify housing units as vacant. Falsification may also occur during listing of a household roster when interviewers deliberately increase or reduce the number of residents in a household, or deliberately and erroneously omit persons with certain characteristics.

As the primary task of the U.S. Decennial Census is to enumerate housing units and their residents, data fabrications are possible at the counting and listing stage. Based on data collected with the Post Enumeration Survey (PES) to the 1990 Census, Tremblay et al. (1991) estimate a total of 179,283 fabricated households. They also identified 420 whole-household fabrications (611 fabricated persons) via quality control reinterviews conducted (by telephone or in-person) with approximately 56,000 PES households. Fifty-nine percent of the 420 fabricated households were initially coded as “complete interview with a household member”, while an additional 32 percent were coded as vacant.

Kindred and Scott (1993) estimated that roughly 0.09 percent, or between 20,000 and 42,000 questionnaires, completed during the Nonresponse Follow-up Operation (NRFU) to the 1990 Census were fabricated. Fabricated questionnaires were identified through a reinterview operation in which a reinterview enumerator would call or visit a housing unit to verify the NRFU enumerator's work, specifically the occupancy status of the unit and the household roster. Higher rates of fabrication were identified among housing units initially classified as non-existent (i.e., not a living quarter), but subsequently found to be occupied or vacant housing units. Relatedly, higher fabrication rates were identified for housing units with 0 persons (coded initially as either vacant or non-existent) or 1 person compared with housing units initially coded as having 6 or more persons. Mulry and Spencer (1991), looking at data from a 1990 census dress rehearsal conducted in Missouri, conclude that whole-household fabrication occurred on between 0.02 percent and 0.23 percent of the sample. And in an examination of a dress rehearsal for the 2000 U.S. Decennial Census, Krejsa, Davis, and Hill (1999) found across three sites that 0 percent to 0.06 percent of interviews were falsified, and that the percentage of interviews for a field interviewer that were reported as partial or proxy interviews and whether a housing unit was vacant were good indicators of whole-household falsification.

2.2.2 Identifying and Contacting the Sample Unit (Interviewer)

Types of falsification that may occur at this stage include:

- Contacting ineligible sample units;
- Assigning incorrect interim disposition codes (e.g., recording a noncontact as a contact);
- Assigning incorrect final disposition codes (e.g., classifying a unit as ineligible [vacant, razed, not a residence] when occupied) (Parsons 2016; AAPOR, 2003); and
- Recording fictitious contact attempts.

Considerably less research has been devoted to possible falsification of paradata or process data (e.g., contact histories of cases and observations of sample units and the larger neighborhood to be completed by interviewers in face-to-face [FTF] surveys). At its simplest, checking to determine whether interviewers are completing their contact histories and observations as required would constitute a simple check on falsification. And many of the detection techniques described here and in later chapters could be applied to process data as well (e.g., duplicate contact history records).

Other examples of falsification of process data might include deliberate under-reporting of contact attempts in surveys that place a cap on calls or deliberately misreporting disposition codes to make productivity appear better than it is (recording a refusal case as ineligible (AAPOR, 2003) or making up contact attempts (Parsons, 2016)).

In a study designed to learn more about interviewers' collection of observational data and paradata, Biemer et al. (2013) surveyed 601 National Survey of Drug Use and Health (NSDUH) interviewers. Among the more interesting findings were that some interviewers reported they would not record a call record if they drove by a housing unit and decided no one was home, in direct violation of protocol, and some reported keeping their cases near the maximum number of call attempts "alive" by not recording additional call attempts. Similarly, Wagner et al. (2017) surveyed interviewers with the National Survey of Family Growth (NSFG) on recording call records (three web surveys: n=29; n=25; n=23). They found that 88 percent of interviewers reported that they would NOT record a call record if they walked past a household and determined that no one was home (in violation of study protocols). An additional 4 percent reported completing their call records at the end of the week at least once, well after contact attempts were made (again, in violation of study protocols). Bates et al. (2010) suggest that under-reporting of attempts resulting in noncontact may be occurring across multiple census-administered surveys. Similar to the Biemer et al. (2013) and Wagner et al. (2017) studies, Bates and colleagues suggest that interviewers may not be recording contact attempts when they drive by a housing unit and determine no one is home. Whether or not the under-reporting is deliberate or simply represents a bias toward recording more salient events (e.g., contact with a sample unit member) was not addressed in their work.

Potential implications of these findings include an upward bias of contact rates, the potential for bias in nonresponse adjustments that employ contact history data, and impacts on estimates of coefficients and predictions in contact models and response propensity models used in responsive or adaptive survey designs (Wagner et al., 2017).

2.2.3 Household Rostering and Screening (Interviewer)

Examples of types of falsification at this stage include screening in ineligible households or screening out eligible households (if screening procedures exist) and including ineligible persons or excluding eligible persons on household rosters. It can include whole or partial roster fabrication, or the roster interview being conducted with an ineligible sample member (Murphy et al., 2016). The motive for interviewing a non-sampled person is to reduce effort required to complete an interview. This can include interviewing a non-sampled willing person in order to reduce effort (AAPOR, 2003) or interviewing someone from a non-sampled unit such as another apartment on the same floor as the sampled unit (Parsons, 2016). In essence, persons can be intentionally added or deleted so that subsequent interviews can be performed with persons present or otherwise more likely to cooperate (AAPOR, 2003).

In the 1997-98 Baltimore STD and Behavior Survey, researchers noticed an inordinately high proportion of cases with rosters containing eligible adult(s) aged 18-45 among a handful of interviewers (Turner et al., 2002). Among interviewers not suspected of falsification, 50 percent of households had at least one eligible adult, whereas among the six interviewers suspected of falsification, 68 percent of households had exactly one eligible adult. For one of those interviewers, all 71 of the households worked had exactly one adult aged 18-45. As a result of this and other collection anomalies, the principal investigator initiated an extensive field verification step in which 100 percent of the cases worked by suspected interviewers were subjected to independent verification, as was 40 percent of the cases worked by non-suspected interviewers. Verification was performed by telephone where possible; otherwise, an experienced interviewer who had not worked on the study attempted verification by personal visit. The research team then reviewed all original interview and verification data for each case. It was determined that the six suspected interviewers had falsified 49 percent of the 451 interviews they conducted.

2.2.4 Conducting the Interview (Interviewer, Field Supervisor)

Much of the data falsification literature has focused on the stage of conducting an interview. In this section, we distinguish between whole interview falsification and partial interview falsification (although the distinction is not always clear). Whole interview falsification is often referred to as curbstoning. At this extreme are cases where an interviewer fabricates the entire interview, never visiting the sample unit. Curbstoning may occur alone or with the assistance of an accomplice (e.g., a friend or neighbor is interviewed). It may involve copying data from an earlier interview (i.e., creating duplicates). Other examples of whole interview falsification are less obvious. For example, the interview may be conducted at the correct sample unit but with the wrong respondent, using a person who is available as a proxy for the selected respondent or as the respondent himself.

In an example of curbstoning or total falsification, Finn and Ranchhod (2017) tested a variety of detection methods using data collected with the second wave of the South African National Income Dynamics Study (NIDS). Of particular use was Benford's law (described in greater detail later in this report). Benford's law posits that the probability distribution of leading digits of continuous measures is not uniform but instead follows a logarithmic distribution (Porras and English, 2004). Its use in detecting data falsification derives from the assertion that possible cheating can be identified by comparing realized distributions of leading digits of responses for each interviewer to the distribution that would be expected if Benford's law holds. Applied to data from the income and expenditure modules of NIDS, interviewers were rank ordered in terms of how far their distributions of leading digits departed from the logarithmic distribution by computing chi-square statistics. The study authors go on to state that four of the five interviewers with the highest chi-square values were later found to have fabricated entire interviews.

Turner et al. (2002) identified interviewer-level anomalies in a survey of the Baltimore population. The study focused on a particularly sensitive topic of sexually transmitted diseases and, not surprisingly, yielded very low participation rates overall. Using interviewer-level cooperation rates, the authors identified six interviewers with success rates between 54 percent and 85 percent; all other interviewers combined achieved a 31 percent cooperation rate. Given the discrepancies, all interviews conducted by these six interviewers were verified through telephone and FTF recontacts. Nearly half of the 451 interviews checked were deemed to be totally falsified.

Koch (1995) explored possible falsification in the 1994 ALLBUS or German General Social Survey. The 1994 ALLBUS used register data as its sample frame, enabling built-in checks of survey data against register data. In comparing interview responses of age and sex to register data, Koch identified 196 of 3,505 interviews as having deviations. Using a mix of postcards, telephone follow-ups, and FTF visits, all 196 discrepant cases were checked. For 45 of those 196, the interview could not be confirmed with sample persons and were, therefore, deemed to be totally falsified. In an additional 51 cases, it was determined that someone other than the selected household member was interviewed. Finally, an additional 31 cases contained erroneous entries to select questions. These cases were retained as interviews, but suspect responses were removed. Hence, Koch (1995) was able to identify cases of both total and partial falsification.

Bredl, Winker, and Kötschau (2012) describe curbstoning in a small-scale survey on land reforms and rural poverty in a former Soviet Union country. Five interviewers (two of whom were study partners) conducted interviews with roughly 200 households in four villages between November 2007 and February 2008. Study directors first became suspicious when an initial set of submitted paper questionnaires showed no signs of dirt or creasing. Narrowing the focus to "clean" questionnaires, the project directors compared responses to multiple questionnaires completed by the same interviewer. Two duplicates were identified, a highly unlikely result given that part of the questionnaire asked for income amounts from multiple sources. The project directors then took a 10 percent subsample of original "clean" interviews and conducted FTF reinterviews. None of the reinterviewed households reported being interviewed.

One form of curbstoning involves duplication or near duplication of observations. Koczela et al. (2015) describe several forms of duplicate cases, including duplicate strings, near-duplicate cases, and full duplicate cases. Duplicate strings of data refer to strings of consecutive responses from one case that appear in one or more other cases in the data file. With near-duplicate cases, a very high percentage of a case's values are duplicated. This scenario may represent a copy-and-paste operation where a whole case is copied and then a handful of item responses is altered. And, in the extreme, are full duplicate cases where, with the exception of case identifiers, an entire case appears more than once in the data file.

Using data collected with the World Values Survey 2005-2008, Blasius and Thiessen (2012) explored within-country duplicate values across 36 widely varying questions in terms of content (gender roles, country goals, financial situation, materialism, technology) and response format (four-point nominal variables, four-point ranking scales, six-point nominal variables, ten-point satisfaction scales). Using multiple classification analysis (MCA), they found that most participating countries had 0 duplicate cases. However, they also identified what they termed to be highly implausible distributions of duplicates in six countries, with one participating country where a quarter of cases had duplicate responses to all 36 items. While the authors concluded that the duplicate responses could easily be explained for one country (possible inadvertent re-entry of data), "…for the remaining countries, the only plausible conclusion is that sample sizes were augmented through copy-and-paste procedures" (65).

Kuriakose and Robbins (2016) assessed the extent to which near duplicate cases were present in major international surveys. In total, they analyzed 1,008 national surveys with over 1.2 million observations. The surveys were collected over a period of 35 years and covered 154 countries, territories, or subregions. The surveys also covered multiple modes of data collection, including computer-assisted personal interviewing (CAPI), computer-assisted telephone interviewing (CATI), and paper-and-pencil interviewing (PAPI). At a minimum, each of the 1,008 surveys included 75 questions. To label a case a duplicate or near duplicate, the authors drew on simulation work that found that surveys including a baseline of 100 questions did not exceed a match rate of 85 percent between two randomly selected observations. Using a publicly available Stata program they developed, *percent match*, and applying the 85 percent match threshold, they found that nearly one in five country-year surveys had a level of near or full duplication of 5 percent or greater. Through additional testing, they found their conclusions to be robust after considering alternative hypotheses for the duplicates, including straight-lining or high levels of item nonresponse, "identical neighbors" given clustered sample designs, and varying numbers of questions in surveys. While the authors never definitively state whether the near duplicate cases they identified represent falsification, they go on to note that the *percent match* program can be used to flag potentially problematic observations, prompting further analysis to definitively determine if they are the result of fraud.

Robbins (2015) used the *percent match* program to identify duplicates in the third wave of the Arab Barometer. Using the country of Kuwait as an example, Robbins identified 178 out of 1,200 cases to have a match rate of 80 percent or greater, while a single interviewer produced a match rate of 80 percent or greater on 122 of 123 interviews conducted. Through further review of this interviewer's workload it was determined that he duplicated his interviews and then changed every 10th response in an effort to avoid identification. Hence, the interviews identified were fully fabricated and appeared as "near duplicates" in the data.

However, caution should be exercised in the use of matching algorithms and similar methods for identifying duplicate or near duplicate records. Through data simulations and application to several real-world surveys, Simmons et al. (2016) identified a number of flawed assumptions in the 85 percent match threshold established by Kuriakose and Robbins (2016). They effectively demonstrated that the percentage of matches identified using the 85 percent match rule varies considerably over survey features, such as number of questions, number of response options, number of respondents, and the homogeneity of the population being surveyed. In their simulation work, for example, Kuriakose and Robbins (2016) assessed only binary variables with a mean of 0.5, a highly unrealistic assumption with regard to most surveys (Simmons et al., 2016). Simmons et al.

(2016) demonstrated that the proportion of respondents that exceed the 85 percent match threshold increases substantially as variable means are allowed to vary across questions. In addition, they found that the percentage of respondent matches increases as the number of questions as well as response options decreases, and the more homogeneous the surveyed population. As the authors note, the use of match methods requires an understanding of "the study-specific environment of a survey to evaluate the meaning of any statistical assessment of the data" (Simmons et al., 2016, p. 335). This work also highlights the importance of using multiple methods for falsification detection.

Partial falsification of an interview involves an interviewer collecting some data legitimately from a respondent and then filling in the remainder of the responses on their own (Blasius and Thiessen, 2012). The faked response patterns may seem plausible to the interviewer given the basic information they legitimately collected and data collected from other similar respondents. The most common form of partial falsification involves collecting a few questions up front, including some basic demographics, and then completing the remainder of the interview at home or some other location. According to Blasius and Thiessen (2012), a rational choice perspective would suggest that an interviewer would collect easily verifiable information, such as age and sex, from the respondent and make up data for more complex content or long, time-consuming batteries of questions. Using a combination of item times, case notes recorded by interviewers, and supervisor consultations with interviewing staff, Dahlhamer et al. (2009) uncovered examples of partial falsification in a large national health survey. The analyses revealed that for a small number of cases the interviewer(s) had spoken with the household and recorded some basic information such as age, sex, whether the household members had health insurance coverage, and any chronic health conditions in an open-ended field for recording notes pertinent to the interview. The interviewer(s) used this information to complete the rest of the interview alone at a later time. In longitudinal surveys, it is common to prepopulate fields in a later wave with data collected in an earlier wave. This usually includes household roster and demographic information. The interviewer could use the prepopulated data to form the basis of their fabrications (although this somewhat blurs the line between full and partial falsification) (Finn and Ranchhod, 2017).

In another example of suspected partial falsification, Dahlhamer et al. (2009) used item times and response data to uncover drum-rolling (the entering of answers, usually the same value, in rapid succession across a set of questions) through five questions on knowledge of heart attack symptoms in a large national health survey. Comparisons of responses to these questions in 2008 versus 2001 revealed unexpected declines in the percentage of adults who understood the symptoms of a heart attack. (All five questions asked about real symptoms of a heart attack. Hence, the correct answer for each question was "yes.") While it was estimated that the five questions could be read very quickly and in their entirety in 20 seconds, the study authors used a more conservative estimate of 13 seconds in their analysis (deemed to be the shortest length of time the questions could be read but still possibly convey their intent). Nearly 19 percent of adult interviews took less than 13 seconds to complete the five items. Furthermore, 42.5 percent of the interviews that took less than 13 seconds to complete these questions had a response of "no" to all five questions. Without separate verification, it was not possible for the study authors to confirm that the interviewers intentionally entered responses of "no" to these five questions (to possibly avoid follow-up questions). Regardless, responses to the symptom questions had to be blanked for several interviews before release of the data.

Finn and Ranchhod (2017) provide an example of partial falsification involving the longitudinal Cape Area Panel Study (CAPS) in South Africa. The fifth wave of data collection included a finger-prick test for HIV status administered by the interviewer. The study expectation was that about 30 percent of women interviewed would be HIV-positive. However, late in the interview period one interviewer returned HIV-positive results for every respondent. Lab results later revealed that the interviewer had submitted blood samples for these cases from the same source. In this particular example, partial falsification was likely due to interviewer sensitivity to requesting and performing the finger pricks.

2.2.5 Finalizing the Case (Interviewer, Field Supervisor)

Research has shown that a common form of falsification involves deliberate miscoding of final case dispositions to make in-scope cases appear to be out of scope. In their work to develop a focused or purposive reinterview program for the Census Bureau, Hood and Bushery (1997) analyzed cases of confirmed falsification and found that more experienced interviewers (5 or more years of experience) falsified a smaller proportion of their workloads than did less experienced interviewers and they tended to do so by classifying eligible units as ineligible. Wetzel (2003) produced similar findings in comparing reinterview results from CAPI and CATI cases for the Consumer Expenditure Quarterly Interview Survey. They found the noninterview misclassification rate to be 4.0 percent for the CAPI cases and 5.6 percent for the CATI cases. Eight of the 19 cases identified as misclassified were occupied housing units originally coded as vacant. Finally, in a Census Bureau falsification study conducted from September 1982 through August 1985, 140 confirmed cases of falsification were identified. While 72 percent of these cases involved complete fabrication of interviews, an additional 17 percent involved misclassifying occupied units as vacant (Biemer and Stokes, 1989).

2.2.6 Recontact/Reinterview (Field Supervisor, Other Field Staff)

We are unaware of any documented cases of falsification at the stage of recontact or reinterview, itself designed to capture, among other assessments of data quality and interviewer performance, instances of potential data falsification. However, it seems likely that the risk of falsification at this stage depends, in part, on who performs recontact or reinterview. That risk increases if supervisors perform the monitoring of their field team(s) (Murphy et al., 2016; Winker, 2016); "…given that a substantial number of classical control procedures such as re-interviews or phone calls are organized through the supervisors, it might be less risky for a supervisor fabricating data than for interviewers" (Winker, 2016, p. 297). The general recommendation is to have recontact or reinterview performed by non-supervisory staff or staff with no involvement in the project.

2.2.7 Data Processing/Aggregation (Data Entry Staff, Research Staff)

Types of falsification at the data processing/aggregation stage include:

- Deliberate mis-entry of values to enhance completeness (e.g., reduce levels of item nonresponse);
- Deliberate miscoding of open-ended, verbatim entries;
- Duplication (or near duplication) of ***entire*** cases to increase the size of datasets, improve the precision of estimates, and/or fit a desired pattern of results;
- Deliberate alteration of respondent-provided responses in an effort to obtain desired findings from the study; and
- Fabrication of whole waves of data in longitudinal surveys using wave 1 or baseline data.

Duplication or near duplication of entire cases is distinguished from imputation, which is a "statistical process that statisticians, survey researchers, and other scientists use to replace data that are missing from a data set due to item nonresponse. Researchers do imputation to improve the accuracy of their data sets" (Rassler et al., 2011, p. 1). In the context of imputation, data are missing for a given item because of a respondent's refusal to provide an answer or because the respondent does not know the answer. In the context of falsification, a researcher or other actor is creating a duplicate or near duplicate of an entire case (no questions are posed to a respondent) in order to increase the number of records in the dataset.

Blasius and Thiessen (2015) used principal components analysis and multiple correspondence analysis to identify statistically improbable response patterns in batteries of items in the 2009 PISA survey. The survey covered 515,958 pupils in 18,641 schools in 73 countries. Principals of participating schools completed the survey. The primary analysis focused on survey questions covering school climate, resource shortages, and management practices. To detect cases of possible partial fabrication, the authors analyzed a string of 184 consecutive items. While single domain items were likely to be correlated, the likely degree of correlation across the entire string was low. Across all 73 countries, 18,019 unique response patterns were identified out of 18,233 cases. Of concern were the 91 duplicates (two surveys with identical responses to the 184 items), 8 triplicates, and 2 quadruplets. The study authors ruled out more benign explanations of the identical response patterns and found evidence in some countries that, in addition to falsification at the respondent level, research institute employees had also duplicated parts of their datasets. Surprisingly, duplicates that were complete with the exception of the school identification number were found in multiple countries.

Koczela et al. (2015) present what they see as evidence of survey falsification during data processing/aggregation via duplication of strings of data. In reviewing survey data, the authors produced a distribution of the frequency of duplicate strings of varying lengths. They found that, on average, two randomly selected cases from the study had a maximum duplicate string length of four to six responses. However, at the very extreme of the tail, they found evidence of complete duplication (i.e., every response to every question [138 total] was duplicated between two or more cases). As a "conservative" threshold, the study authors used a cutoff of 50 percent of duplicate responses (69 or more responses were exactly the same between 2 or more cases) to identify cases for more intensive investigation. While they present evidence of falsification at the interviewer and data entry staff level, they also found strong clustering of lengthy duplicate strings at the supervisor level.

Falsification at this stage may grow over time. Koczela et al. (2015) discuss “machine-assisted” fabrication in which computers can be used to duplicate cases or, based on more sophisticated algorithms, generate “plausible” records, with the intent to fabricate, as distinct from statistical imputation.

3. Preventing Falsification

3.1 Introduction

Survey organizations prevent data falsification by creating a workplace culture that limits falsification, selecting study design features that help minimize the risk of falsification, and ensuring that best practices are being followed for training and monitoring of interviewers and research staff before, during, and after data collection. Preventing data falsification is fundamentally important for research organizations to protect the quality of survey data. These measures are also critical to maintaining the goodwill of survey participants and ensuring ethical research conduct. In this chapter, we examine how research organizations evaluate and minimize incentives for data falsification through measures conducted before, during, and after a study with the eye toward preventing falsification. Given the difficulty of and general lack of empirical evidence evaluating and comparing preventative measures, the task force cannot identify the most or least effective measures for prevention. However, we do recommend best practices where available.

Cressey (1953) first posited a "fraud triangle" model for white-collar financial crime that can apply to issues of falsification. Cressey examined the motivations to commit fraud in different contexts and identified the three points (or sides) of the triangle that consist of:

- Pressure or motivation to commit the act;
- Perceived opportunity; and
- Rationalization (Wells, 2017).

Pressure or motivation to commit the act of fraud occurs when the individual feels that problems (e.g., financial) cannot be shared with others in the organization, leading this individual to resolve these problems using devious means (Dellaportas, 2013). The opportunity arises when the perceived chance of being caught is low. When a person rationalizes the act of fraud, this reduces their cognitive dissonance over the potential problem of committing the crime (Dorminey, Fleming, Kranacher, and Riley, 2012, p. 558; Wells, 2017, p. 7-8). To the extent possible, issues of pressure, opportunity, and rationalization that have implications for interviewers to potentially falsify data should be considered during the study design phase. The potential to falsify may be reduced if these factors can be minimized or procedures put in place to appropriately train interviewers, monitor their work, and evaluate the survey data and paradata for the presence of suspected falsification.

3.2 Organizational Factors

The organization sets the tone for all interviewers and research staff as to what is acceptable or unacceptable work behavior. This may be reflected in the somewhat amorphous term of "organizational culture" or in policy documents and practices at the institutional level, such as required human subjects training. Following Cressey's fraud triangle, organizational factors can help reduce the pressure or motivation to falsify data, mitigate the opportunity to do so, and make it difficult for the individual to rationalize the act of falsification. For instance, Dorminey et al. (2012) report that employee dissatisfaction and poor working conditions lead to more employee fraud. Thus, organizational factors are key to mitigating data falsification.

3.2.1 Organizational Culture

Institutional leadership is critical for creating an organizational culture that values research integrity, and, as such, seeks to prevent data falsification and other forms of scientific misconduct (Institute of Medicine [IOM] and National Research Council [NRC], 2002). Barney (1986, p. 657) defines organizational culture as "a complex set of values, beliefs, assumptions, and symbols that define the way in which a firm conducts its business." Applying this idea to survey organizations, Kennickell (2015) argues that survey organizational culture is fundamental for preventing interviewer curbstoning and maintaining confidentiality, especially in FTF surveys where monitoring is difficult. To create an organizational culture that minimizes the risk of falsification, organizational leaders (managers, field supervisors) must clearly articulate the survey organization's "values, goals, and rules" (AAPOR, 2003). This applies to all levels of a survey organization but may be especially important for employees who receive little day-to-day supervision or interaction.

As indicated in a 2017 National Academies of Sciences, Engineering, and Medicine report on research integrity, exactly how the research environment affects decisions by researchers to engage in scientific misconduct is not clear, with many, varied hypotheses. In general, individuals are considered to have their own tendencies to potentially engage in data falsification or fabrication, and the environment in which they operate may heighten or lower the potential for them to do so. For example, in a review of reported reasons in research misconduct cases investigated by the ORI, Davis, Riske-Morris, and Diaz (2007) found that organizational factors such as insufficient supervision and non-collegial work environments led to research misconduct. But, in this case study, the researchers did not evaluate whether these cited issues occur at the same rate in research situations that do not result in research misconduct. Empirical research on whether workplace culture prevents data falsification in survey organizations and in research more generally is even more sparse. For instance, Koczela, Furlong, McCarthy, and Mushtaq (2015; p. 420) argue, "Survey organizations have to cultivate the idea that interviewers are on the same side as headquarters personnel and survey project leaders," and that these individuals, in turn, are all working together to produce data that will ultimately benefit the respondents. No empirical evidence is provided to demonstrate how to do this or whether this approach is efficacious. Although issues related to workplace culture and how information is transmitted from manager or supervisor to employees is often cited as key for creating a climate devoted to ethical research (e.g., DuBois and Antes, 2018), actually evaluating the association between the research environment and data falsification or fabrication is difficult (IOM and NRC, 2002). Randomly assigning organizations to high versus low "ethical research environments" is impossible; however, conducting observational studies that ask interviewers or researchers to evaluate the research environment at their organization and correlating these responses with observed data on falsification may be more fruitful.

Thus, although there is consensus that environmental and workplace cultural factors are critically important for preventing data falsification and fabrication, there is little to no empirical evidence on what workplace cultural factors are relevant, efficacious, or most important. Despite this lack of evidence, best practices include regularly communicating norms and expectations about the research process, research misconduct, and each researcher's individual rights and responsibilities for maintaining this process, as well as evaluating the research culture directly (National Academies of Sciences, Engineering, and Medicine, 2017). This suggests that ***survey organizations should regularly reaffirm commitment to the norms and expectations of researchers, interviewers, and field supervisors to 1) collect accurate and reliable data from the selected households and respondents; 2) ask each question in a survey instrument; and 3) ensure that the interviewers and field supervisors are personally responsible for maintaining confidentiality, for visiting the selected housing units, and for reporting any problems that they encounter immediately to the supervisory and managerial staff.*** It also suggests that this training should not come solely at the beginning of a data collection effort but, to create a workplace culture that minimizes the risk of data falsification, that ***research ethics and the importance of not falsifying data should be communicated intermittently throughout data collection.*** During the hiring and training phases, research managers and supervisory personnel should communicate to interviewers and research staff a solid understanding of survey methods, the nature and purpose of the research, and how the collected data will be used. This promotes buy-in to the research and emphasizes the important role that interviewers play in its success (Parsons, 2016). Finally, survey organizations should consider turning the survey lens inward, ***conducting internal surveys on the climate of research integrity at all levels.***

3.2.2 Research Ethics and Other Human Subjects Training

The cornerstone of most institutions' approach to mitigating data falsification is research ethics training. Organizations should train interviewers about data falsification, why falsification is problematic, and how to avoid it (e.g., Johnson, Parker, and Clements 2001; Murphy et al., 2016; Thissen and Myers, 2016). As discussed in Chapter 6, most SROs responding to our survey reported requiring some sort of research ethics training of their interviewers and research staff. For example, the GESIS CAPI interviewer training (Daikeler, Silber, Bosnjak, Zabal, and Martin, 2017) includes a 3-plus hour module on "Professional Standards and Ethics, Data Protection and Privacy." In many organizations, research ethics training generally occurs in the form of online modules on research compliance (DuBois and Antes, 2018). Across different ethical trainings, basic guidelines about ethics, codes of conduct, the common rule, protection of human subjects, professionalism, scientific misconduct, privacy and confidentiality, and conflicts of interest are among the topics covered in responsible conduct of research trainings (Watts et al., 2017).

To our knowledge, the efficacy of human subjects training in preventing data fabrication and falsification among survey interviewers has not been directly evaluated. In a recent meta-analysis of the efficacy of responsible conduct of research (RCR) ethics programs across fields and types of training, Watts et al. (2017) found that RCR training did improve knowledge of ethical issues and the participants' perceptions of their knowledge and ethical attitudes. However, this was across multiple fields and distinct from the kinds of outcomes of most interest to survey organizations – prevention of data falsification. Although training seems intuitively necessary, what constitutes the best practices for training to reduce the risk of data falsification among interviewers is unknown. To our knowledge no test of the effectiveness of alternative training practices has been published or reported.

3.3 Preventing Falsification: Study Design Factors

Beyond the organizational factors that may be instituted to reduce the likelihood of falsification by data collection staff, study-level factors play an important role. While literature citing quantitative evidence of the association between study design and the potential for falsification is limited, several sources discuss intrinsic relationships based on the nature of design elements and their likelihood to invite or prevent falsification. Under Cressey's fraud triangle, study design factors affect the pressure that an interviewer may feel to fabricate or falsify data, as well as the opportunity to do so.

Study design elements that may be related to the tendency to falsify were first documented by Crespi (1945). He identified the questionnaire itself as a tool that could invite or prevent falsification. Specifically, Crespi notes that excessive questionnaire length and burden could lead to situations where the interviewer is tempted to falsify the interview rather than put forth the effort to gain the cooperation of the respondent and administer the survey. Additionally, repetitive, difficult, or antagonistic survey questions could lead to the situation where "fabrication is made to appear the only practicable solution to the problems facing the interviewer" (p. 436).

Josten and Trappman (2016) identified the presence of filter questions as a potential temptation for interviewers to manipulate answers and minimize the number of follow-up questions. They found that interviewers minimizing the number of follow-up questions to occur more frequently among field interviewers who are paid by case and not closely supervised. They did not see this behavior with telephone interviewers who were paid hourly and closely monitored. The authors also noted that questionnaire sections with elaborate or involved logic and skip patterns, such as social networks, raise the potential for partial falsification. Studies with additional burdensome requirements, such as sensitive questions or biomarker collection, may also lead to higher potential falsification (Turner, Gribble, Al-Tayyib, and Chromy, 2002).

The presence of respondent incentives may also play a factor in interviewers' motivation to commit falsification (Murphy, Baxter, Eyerman, Cunningham, and Kennet, 2004; Murphy et al., 2005). While this has not been experimentally tested, the presence of large, cash incentives may tempt interviewers to fraudulently complete the interviews without the respondent present in order to keep the cash for themselves

To prevent falsification, study-level factors should be taken into account and addressed. To summarize, the above references identify these factors to be associated with a higher likelihood of falsification:

- Using long and burdensome questions and questionnaires;
- Using complex survey logic and study protocols;
- Relying on survey modes that are more difficult to monitor (e.g., field vs. telephone);
- Clustering of the target population in areas with low resources or chronic difficulty with completing interviews; and
- Presenting large cash incentives or other materials that may create a temptation for the interviewer.

The study design factors specific to surveys mostly fit into the first two categories of Cressey's fraud triangle (pressure and opportunity), though rationalization may play a role if interviewers are not adequately trained to understand the implications and consequences of committing falsification.

3.4 Preventing Falsification: Personnel Factors

Preventing falsification may be related to the structure of how interviewers are hired and paid and the areas in which they work. In this section, we review personnel factors that may prevent (or contribute to) data falsification. These personnel factors may act primarily on the pressure/motivation and rationalization parts of the Cressey Fraud Triangle by creating situations where interviewers feel motivated to complete their interviews quickly rather than well, and perhaps allow them to rationalize their behavior.

3.4.1 Hiring Practices

Different hiring practices and selection criteria have been hypothesized to prevent data falsification and fabrication (Boyd and Westfall, 1955). For instance, Biemer and Stokes (1989), citing Crespi (1945), suggest that part-time interviewers may be more likely to fabricate interviews because of competing time demands with other jobs but provide no empirical data about this hypothesis.

3.4.2 Outside Stressors

Less-than-ideal interviewing conditions may contribute to interviewers fabricating or falsifying data. If interviewers find it particularly difficult to contact householders or gain cooperation, they may consider it easier to make up data than to continue trying to gain cooperation. For instance, interviewers who are located in neighborhoods that are "difficult to interview," which may include urban areas, have been hypothesized and, in some large national surveys, found to be more likely to falsify interviews (Crespi, 1945; Biemer and Stokes, 1989; Breiman, 1994; Gwartney, 2013; Shraepler and Wagner, 2005). For instance, Biemer and Stokes (1989) note that the Current Population Survey and National Crime Survey both found falsification rates to be significantly higher in urban areas compared with rural areas. This may be related to increased difficulty to contact and gain cooperation with respondents in urban areas, possibly due to the presence of crime and building "gatekeepers," suggesting that studies with a largely urban sample may require significant resources devoted to preventing falsification. This may also hold true at the country level, where low- and middle-income countries may employ survey teams having limited supervision and facing chaotic environments, language barriers, and low literacy. Such factors can be associated with a higher likelihood to falsify (Birnbaum, DeRenzi, Flaxman, and Lesh, 2012). Additionally, bad weather or other externalities have been posited as being a cause of fabrication (Crespi, 1945; Biemer and Stokes, 1989), but again, empirical evidence is lacking.

Other geographic factors of a study may also impact the likelihood for falsification if cases are highly clustered in areas where few professional interviewers are available. In such cases, studies may need to hire inexperienced staff, who may be more likely to commit falsification than experienced survey interviewers (Turner et al., 2002).

3.4.3 Interviewer Pay and Pay Structure

Poor interviewer pay and a per-interview (or bonus for completed interview) pay structure are hypothesized to be reasons why interviewers fabricate data (Koczela et al., 2015; Crespi, 1945; Biemer and Stokes 1989; Valente, Dougherty, and Stammer, 2017). As such, better pay and an hourly or salaried pay structure are hypothesized to prevent fabrication (Bennett, 1948; Bredl, Storfinger, and Menold, 2011; Smith, MacQuarrie, Herbert, Cairns, and Begley, 2004; Turner et al., 2002; Winker, 2016). Empirically, interviewers in Sub-Saharan Africa cited poor pay as a reason that they may fabricate data (Kingori and Gerrets, 2016). However, Kreuter, McCulloch, Presser, and Tourangeau (2011) found no difference between CATI interviewers who were paid per completed interview compared to those paid hourly in completion of filter questions, an issue of partial falsification. In FTF surveys in Germany, on the other hand, Josten and Trappmann (2016) found evidence of CAPI interviewers who were paid per interview obtained lower numbers of names in a series of questions about the respondent's social network than CATI interviewers on the same study who were paid by the hour (see also Kosyakova, Skopek, and Eckman, 2015). In a lab experiment, Landrock (2017) randomly assigned interviewers to be paid per hour versus per complete. Then, the interviewers were asked to complete both real interviews and to intentionally falsify interviews. The study found that the payment scheme had no effect on the real data but was associated with non-differentiation in the falsified data.

3.4.4 Length of Employment

Less experienced interviewers are hypothesized to be more likely to falsify data, and the empirical evidence largely supports this hypothesis (Biemer and Stokes, 1989; Schraepler and Wagner, 2005; Turner et al., 2002; Hood and Bushery, 1997). Study-specific experience appears to be relevant – Schraepler and Wagner (2005) found that the interviewers who falsified interviews were new to interviewing on the German Socio-Economic Panel, even though they had been interviewers as a career for longer. When more experienced interviewers falsify data, they tend to falsify only part of an interview rather than complete interviews (Turner et al., 2002; Bredl et al., 2011) or to reclassify sample units as ineligible rather than making up data for a full interview (Hood and Bushery, 1997). One possible way to ameliorate this is to assign new interviewers to surveys with less complicated designs and, of course, monitor these interviewers more aggressively as they develop their skills.

3.5 Preventing Falsification: Methods Used Before Data Collection

Methods used prior to data collection to prevent falsification focus on training staff about the consequences of falsifying data, making it difficult to rationalize that falsification is permitted as part of the job. This includes emphases on consequences for the falsifier, the organization, and the scientific integrity of the data. Previously, it was recommended that organizations train interviewers about falsification, but unlike human subjects' protection training, this training is not mandatory. In some organizations, leadership might assume falsification is covered in the required human subjects protection training, but depending on how the organization fulfills that requirement for its staff, that may not be the case. Many of the typical ways organizations do this is by requiring interviewers take the online NIH training, the Collaborative Institutional Training Initiative (CITI) Social and Behavioral Research Training module or some portion of it, or developing an in-house training and testing program (AAPOR, 2009). Aside from in-house programs, which presumably vary widely, the other methods are heavily focused on the history of research ethics, assessing risk, informed consent, and privacy and confidentiality – not falsification and its consequences.

Thus, specific training about falsification should be a best practice. In addition to training the staff about falsification and its consequences, an organization may want to require its staff to sign a pledge of ethical behavior or a "data integrity" agreement. Just as many organizations require staff to sign a confidentiality agreement as a condition of employment, a data integrity agreement follows the same structure. The pledge should describe the importance of data integrity, outline what constitutes falsification and fabrication, and note the consequences to the interviewer and other staff in the event falsification occurs. (Johnson, Parker, and Clements, 2001; ORI, 2002; Parsons, 2016). Project-specific training should reinforce the importance of data integrity and the consequences of falsification. As part of this training, interviewers should be provided with clear definitions of data falsification. Depending on the project, this may include final case dispositions applied to never contacted cases or purposively coding eligible cases as ineligible, entering responses to questions that were not asked (for a part of or the entire interview), interviewing the wrong respondent to save time, completing a CAPI interview without using the laptop, and entering false responses to gate questions to avoid detailed follow-up questions (Johnson et al., 2001; Murphy et al., 2016).

Of course, adequately training staff to prevent falsification assumes the principal investigators and other organizational leaders are themselves adequately trained to lead on this issue. DuBois, Chibnall, Tait, and Vander Wal (2016) found that among researchers referred for research compliance or integrity remediation training, the most common reason was not willful malfeasance but failure to provide adequate oversight of their projects. Antes, Mart, and DuBois (2016) found that while principal investigators thought leadership and management "are essential to performing effective research," they felt their scientific training did not adequately prepare them for this. Thus, preventing falsification may require that organization leaders are adequately trained to lead and oversee their projects and create a culture that discourages falsification.

3.6 Preventing Falsification: Methods Used During Data Collection That May Be a Deterrent

Preventing falsification during data collection focuses on reducing the (perceived) opportunity to falsify data without being caught. Thus, many of the methods that are used for preventing falsification are based on deterring falsification. As a group, these methods, including monitoring, use of CARI, analysis of paradata, use of verification (re)interviews, and real-time use of GPS devices, have received little attention or empirical evaluation of their efficacy as a deterrent to falsification.

3.6.1 Monitoring Interviews

Monitoring interviewers during data collection is an important deterrent to falsification. As Thissen and Myers (2016) note, being watched has a powerful effect on people's behavior. In CATI facilities, it is common practice to use "silent monitoring" (audio, video, screen capture) to ensure compliance with interviewing procedures (AAPOR, 2003; Harding and Jackson, 2012). Monitoring is usually conducted by a supervisor who listens to or observes a portion of the interview, takes notes, and provides feedback to the interviewer. Standard practice is to have the interviewer inform the respondent at the beginning of the call that a supervisor may listen in to ensure that the interview is being conducted properly, with between 5 and 15 percent of cases monitored (Biemer and Lyberg, 2003; Murphy et al., 2016). Monitoring alone is usually thought to be sufficient for detection and deterrence of falsification in these settings (AAPOR, 2003). In FTF surveys, supervisors observe new interviewers for a limited period of time, usually their first full field assignment, to ensure compliance with established interviewing procedures. If performance problems arise over time, interviewers may also be "placed into" supplemental observation.

3.6.2 CARI

CARI is an additional monitoring tool that may serve a deterrent role.[1] Although not in real time, CARI enables a similar level of monitoring in-person surveys to that traditionally available in CATI facilities. CARI involves the digital recording of interviewer-respondent interactions during an

[1] However, the extent to which CARI acts as a deterrent is unknown. McGonagle, Brown, and Shoeni (2015) found that recording interviews improved data quality at a tradeoff of longer interviews.

interview. It is relatively unobtrusive, making use of the microphone and soundcard on an interviewer's laptop or mobile device (Keating, Loftis, McMichael, and Ridenhour, 2014). While the entire interview may be recorded, a more common approach is to enable audio recordings at pre-specified questions or a random sample of questions. At the outset of the interview, the respondent is informed that a portion of the interview may be recorded for quality control purposes. The respondent is also assured that the recordings will not affect the confidentiality of their responses. If consent is given, the recording mechanism is started. Like other aspects of the interviewing process, the respondent reserves the right to stop that recording at any time. As with the survey data, the audio recordings are transmitted over the Internet via a secure, encrypted connection to the survey organization.

An important use of CARI is for interview verification, especially in FTF surveys (Biemer, Herget, Morton, and Willis, 2000). To detect falsification, CARI coders or monitors listen for several potentially problematic circumstances on the audio recordings: no voices on the audio file while clicking or room noises are audible; an interviewer's voice may be heard but no respondent voice is captured on the recording; the respondent answers too quickly or laughs in inappropriate places; the same respondent's voice is heard in recordings of multiple interviews (Thissen and Meyers, 2016; Thissen and Rodriguez, 2004). While CARI is an important falsification detection tool, it also may serve as a deterrent. Interviewers need to be trained on the purpose and use of CARI. As such, it should be immediately clear that CARI will be used to monitor their work and ensure interviewing procedures are followed (Smith and Sokolowski, 2008). In their initial feasibility testing of CARI, Biemer et al. (2000) performed interviewer debriefings to ascertain their feedback and concerns about CARI. Roughly 87 percent were positive or neutral about the use of CARI for detecting falsification.

3.6.3 Paradata Analyses

Analysis of paradata during data collection can also be used to detect and potentially deter interviewer falsification. For example, Murphy et al. (2016) describe real-time monitoring with the use of survey dashboards examining length of interview, along with other metrics that can detect and, perhaps, prevent falsification. The U.S. Census Bureau's Performance and Data Analysis tool (PANDA), initially developed for the American Housing Survey and later adopted by surveys such as the National Health Interview Survey, is a web browser-based tool that summarizes and presents collected survey data and paradata, the latter captured via interviewer observations and audit trails (i.e., files recording keystrokes captured by the CAPI instrument) (Jans, Sirkis, Schulheis, Gindi, and Dahlhamer, 2011; Stringer and Dahlhamer, 2010). From a deterrence perspective, the use of paradata as a monitoring and quality control tool is presented to interviewers in training; thus, as with CARI, paradata may be helpful in preventing falsification since interviewers understand their performance is being routinely monitored.

3.6.4 Verification Interviews/Reinterviews

A common method for detecting and potentially deterring falsification in FTF surveys is the verification interview or reinterview (Biemer and Stokes, 1989; Forsman and Schreiner, 1991). Reinterview involves recontacting original survey participants to briefly verify that the correct sample units were visited, the interview took place, whether the outcome or disposition of the original interview was recorded correctly, and all questions were asked, and to ensure that proper procedures were followed (e.g., the interviewer used his/her laptop computer to complete the CAPI interview) (Dajani and Marquette, 2015; Murphy et al., 2016). From a deterrence perspective, interviewers should be informed that a part of their workload will appear in reinterview at any given time during data collection.

3.6.5 GPS

For FTF interviews, GPS data have recently emerged as a tool for detecting and deterring falsification (Edwards, Maitland, and Connor, 2017; Ellis, Sikes, Sage, Eyerman, and Burke, 2011; Eng et al., 2007; Thissen and Myers, 2016). GPS coordinates can be captured actively by the interviewer or passively by software during survey administration (Keating et al., 2014; Thissen and Myers, 2016). The captured GPS coordinates are tagged (geotagging) to the sample address listing and/or survey responses and algorithms are used to compute distances between the actual GPS readings and the expected coordinates. The GPS data, therefore, verify that interviewers were at a particular location at a particular time (Edwards et al. 2017; Ellis et al., 2011; Eng et al., 2007; Thissen and Myers, 2016). Of significance from a deterrence perspective, especially with active capture of GPS coordinates, is that interviewers are made aware that 100 percent of their workload is being monitored in near real time.

A GPS-based tool with significant deterrence capabilities is geofencing. With geofencing, survey managers can set virtual perimeters or geospatial boundaries around sample units, alerting interviewers if they appear to be outside those boundaries or too far from a sample unit at the start of an interview (Keating et al., 2014). Alerts or warnings can be sent to interviewer cell phones or other mobile devices (Burke, 2015; Dajani and Marquette, 2015).

3.7 Preventing Falsification: Methods Used as Consequences That May Be a Deterrent

Methods used during data collection to monitor interviewer performance should be mentioned during interviewer training. Interviewers need to know they will be monitored, just not *when* they will be monitored (AAPOR, 2003). Careful supervision of interviewer and research staff is critical to prevent data falsification. Generally speaking, when interviewing and research staff receive proactive feedback and formal support from their supervisors, they are less likely to falsify survey data (Johnson et al., 2001). Additional suggestions in the literature include hiring interviewers who can work fairly regular shifts (making monitoring easier) and conducting reference checks (and possibly criminal background checks) (Parsons, 2016).

The consequences that a person suffers if it is determined that they have falsified data may serve as a deterrent for others, thus, making it more difficult to rationalize falsification. Knowing about consequences may also make the reconciliation process more challenging, increasing the cognitive dissonance that an individual may feel about falsifying data (Dorminey et al., 2012). Personnel actions, from reprimands to dismissal, have been and should be initiated if it is determined that an individual has falsified data (AAPOR, 2003; Fanelli, 2009). For interviewers and their supervisors who are identified as having falsified data, the loss of respect and subsequent treatment by their colleagues can be harsh. These personal and public consequences can also serve as deterrents to data falsification.

4. Detecting Falsification

4.1 The Need for and Challenges of Detection

While prevention is the first and best solution to mitigate the potential for falsification, varied incentive structures make it difficult to foresee all threats to data integrity. As a result, it is important to have detection processes in place; proper documentation of the detection process and ultimate impact on data quality is essential to maximize effectiveness (Winker, 2016).

Entire interviews that are fabricated may be easier to detect than individual items. The latter requires more extensive verification, which can result in a tradeoff between the cost of verification and a greater probability of detecting falsification (AAPOR, 2003).

Differences in the patterns with which an interviewer or supervisor may falsify increase the complexity of detection efforts. Falsification may be *random* in that there is no clear pattern to which cases or items are fabricated. Or it may be done in a *purposive* fashion based on the type of case, item, or period in data collection. For instance, an interviewer may falsify those cases where an initial noncontact or refusal prevents easy completion. They may falsify sections of a survey that deal with sensitive topics or those requiring burdensome access of records. Or, facing performance goal pressures, an interviewer may be tempted to falsify toward the end of a data collection period to meet certain prescribed goals. Whether falsification is conducted in a random or purposive fashion, it may be a frequent or infrequent event. When an interviewer falsifies a larger amount of his/her work, it may be easier to detect if the patterns observed differ from what would be expected from valid interviews. Infrequent falsification may be more difficult to detect since fraudulent cases combined with a large number of valid interviews may mask patterns suggestive of falsified data.

Finally, interviewers or supervisors who commit falsification may use strategies with different levels of sophistication. At the simpler end of the spectrum, interviewers may take little effort in covering up their behavior, falsifying interviews at odd times of the day, in very little time, in large batches, or using answers that may not reflect what one would expect from a valid interview. If audio monitoring is present, a simple falsification strategy may not attempt to mimic the voice of a valid respondent. More savvy actors may take extra steps to understand the checks in place and devise complex strategies to remain undetected (Murphy et al., 2016).

4.2 Methods/Technology Available for Detection

Just as the types of falsification are varied, the methods available to detect possible falsification during data collection include a wide range of approaches. Several of these methods (interviewer monitoring, the use of CARI, paradata analysis, the verification interview (or reinterview), and the use of GPS methods including geofencing) are discussed in section 3.6, in the context of the deterrent effects of alerting data collection staff to the use of detection methods. In this section, more detail is given on the use of these technologies, and case studies are cited as examples of their application.

When considering and applying methods of falsification detection, it is important to bear in mind that if a data point or pattern does not fit the researcher's expectation based on knowledge of the survey, respondents, or the subject matter, falsification may be one of several possible explanations. Other explanations could include measurement error due to reasons other than falsification, including unintentional interviewer errors, respondent confusion, satisficing behavior, etc. Systems that include regular detection activities of one or more phases or types can be said to have quality assurance (QA) practices in place, which Dajani and Marquette (2015) define as "the planned and systematic actions necessary to provide adequate confidence that a process meets proposed quality standards" (p. 4).

4.2.1 Methods That Involve Review of the Interview "Process"

Several methods of falsification detection involve monitoring or validating the interview process. It is common practice in centralized telephone facilities to use "silent monitoring" (audio, video, and/or screen capture) to ensure compliance with interviewing protocols (AAPOR, 2003; Harding and Jackson, 2012). The monitoring is usually conducted by a supervisor who listens to/observes a portion of the interview, takes notes, and provides feedback to the interviewer. It is common practice to have the interviewer inform the respondent at the beginning of the call that a supervisor may listen in to ensure that the interview is being conducted properly (Biemer and Lyberg, 2003). Monitoring alone is usually sufficient for detection and deterrence of falsification in these settings (AAPOR, 2003). In FTF surveys, new interviewers are observed by supervisors for a limited period of time, usually on their first full field assignment, to ensure compliance with established interviewing procedures. If performance problems arise over time, interviewers may be "placed into" observation.

A more common method for detecting falsification in FTF surveys is the verification interview or reinterview (RI). The U.S. Census Bureau (Census Bureau), for example, has a long history of using RI for conducting reliability (test-retest) analyses and detection of falsification (Bushery, Reichert, Albright, and Rossiter, 1999). Schreiner, Pennie, and Newbrough (1988) used data collected with Census Bureau-administered surveys from 1982 to 1987 to highlight the importance of the verification interview as a means of detecting fraud. Among their findings: 83 percent of suspected falsifications were confirmed via RI; the majority of falsification involved total rather than partial fabrication of respondent-level data; falsification rates ranged from 0.4 percent to 6.5 percent depending on the survey; and interviewers with more experience were significantly less likely to falsify.

RI is often completed via a mixed-mode design, starting with telephone recontact (to minimize costs) and then moving to FTF if necessary. (Cases where no telephone number was provided or certain types of dispositions [e.g., building demolished] may necessitate personal visits from the outset.)[2] Depending on the survey organization and/or the requirements of the particular study, roughly 5 percent to 15 percent of the original sample may be recontacted for verification purposes (AAPOR, 2003; Murphy, 2016). Often, a two-pronged approach for selecting cases is used, whereby one prong involves a random sample of cases (usually a multistage design involving stratified, systematic sampling of interviewers followed by a random sample of cases from a selected interviewer's workload) and the other involves purposive or focused selection of cases, (i.e., cases are selected due to anomalies identified in the data or in the interviewer's performance, including previous history). To minimize recall error, recontact or RI occurs as soon after the original interview as feasible. The Census Bureau, for example, has a stated goal of recontacting a sample unit within 2 weeks of final dispositioning of the case (Dajani and Marquette, 2015).

A number of limitations to the use of RI for detecting falsification have been noted. For one, it is commonly applied to a small subset of all cases. For example, approximately 2 percent of Current Population Survey (CPS) interviews are sampled for RI each month (Li, Brick, Tran, and Singer,

[2] A potential indicator of falsification is a high rate of missing telephone numbers for an interviewer's workload. Non-collection of telephone numbers may be deliberate so as to circumvent RI and other quality control procedures.

2011). Hence, the full scope of falsification may go undetected. Second, if falsification (or RI itself) is taking place at the supervisor (or higher) level, several instances of falsification may go unnoticed or unreported.[3] Third, respondent recall may be problematic, leading to both false positives and false negatives (Bredl, Storfinger, and Menold, 2011; Menold and Kemper, 2014). In an early study of recontact, for example, Hauck (1969) found that 14 of 100 non-interviewed persons actually reported being interviewed. Fourth, to minimize respondent burden, the RI is relatively short. Therefore, the amount of information from the original interview that can be verified is limited, meaning some cases of partial falsification may be missed (Menold and Kemper, 2014). Fifth, like the original interview, there is also nonresponse to the RI (Bredl et al., 2011). Finally, RI is expensive (Murphy et al., 2016).

Given the need for efficiency due to cost constraints and other limitations, a sizeable body of research has focused on the benefits of purposive or focused RI. Li et al. (2011), for example, demonstrated the utility of using statistical modeling for informing RI sampling methodologies with the CPS. At the time of the research, the standard CPS approach to selecting cases for RI involved a two-stage stratified sample design over a 15-month cycle of interviews. Sampling at the interviewer level was performed in such a way as to ensure that an interviewer was in RI between one and four times within a 15-month cycle. Interviewers with less than 5 years of experience on CPS were sampled more frequently than experienced interviewers. Paradata-based measures (e.g., interview length, month in sample) and interviewer experience were included as covariates in a logistic regression of falsification status (confirmed falsification or "still suspected" versus not falsified). Using the predicted probabilities from the model, cases were rank-ordered on the likelihood of falsification and three alternative sampling designs were proposed: proportional to size (PPS), truncated, and stratified. Applied to real and simulated data, it was found that all three alternative sample designs would yield more cases of falsification than the traditional design (from roughly 1.7

[3] An important consideration with this approach is who conducts the RIs. For example, some survey organizations have interviewer supervisors conduct RIs. This is often predicated on the assumption that a supervisor will have unique knowledge of the areas in which their interviewers work, and that area characteristics may aid in explaining suspicious response patterns in the data or interviewer outliers. As Winker (2016) noted "…given that a substantial number of classical control procedures such as re-interviews or phone calls are organized through the supervisors, it might be less risky for a supervisor fabricating data than for an interviewer" (p. 297). In response, many research organizations conduct RI out of a centralized location, such as a CATI facility.

times more falsified cases in the PPS design to 2.3 times more falsified cases in the truncated design).

Hood and Bushery (1997) examined survey responses and metadata on case dispositions in their work to develop a focused RI program for the National Health Interview Survey (NHIS). The authors relied on a basic assumption that cheaters would attempt to "keep it simple" (p. 820). More specifically, they assumed that cheating interviewers would opt to code eligible cases as ineligible or screen out minority households as nonminority to avoid a longer interview. To test these assumptions, they developed interviewer-level indicators of the ineligible unit rate, nonminority screening rate, short interview rate, and the no telephone number rate. To account for the spatial homogeneity of an interviewer's assignment area, 1990 census data were used to create a set of comparable measures at the county or segment level. For example, the proportion of units with no telephone was generated for each county or segment within a county. The difference between an interviewer's rate and the expected rate for the assignment area, based on census data, was then computed. As a next step, standardized scores for each of the four difference measures were computed and summed. Interviewers with scores greater than 1.5 were flagged as multivariate outliers and placed into focused RI. Preliminary results were promising. Focused RI detected three falsifiers among a set of 83 interviewers checked, yielding a 3.6 percent hit rate compared to the 0.2 percent achieved by random RI.

As a further example, Krejsa, Davis, and Hill (1999) developed a set of interviewer-level metadata indicators, such as the proportion of vacant households in a caseload or the proportion of cases completed at the end of the interview period, to identify outlying interviewers. These interviewers were subsequently placed into focused RI. Whereas random RI identified one falsifying interviewer based on 1,706 cases checked, the focused RI approach identified 10 falsifiers based on 1,737 cases checked. Finally, Bushery et al. (1999) describe the use of paradata with the NHIS to aid in identifying interviewers with performance issues or suspicious outcomes. Time and date-based metrics found to be promising included the number of interviews completed in a single day, the number of interviews completed in the final 2 days of the interview period, and the number of cases finalized as ineligible in the final 2 days. Using statistical process control (SPC) methods, the authors flagged outlying (outside three standard deviations from the mean) interviewers and recommended them for purposive or focused RI.

Technology advances have enabled more sophisticated methods of falsification detection, including CARI, image or screen captures, and geotagging. CARI is a particularly useful tool for monitoring CAPI interviews and has largely been used for detecting and reducing measurement error (Edwards and Maitland, 2015; Hicks et al., 2010). It is relatively inobtrusive and occurs directly through the laptop or a mobile device (Keating, Loftis, McMichael, and Ridenhour, 2014). Consent to record is obtained by the interviewer at the outset of the interview. While entire interviews may be recorded, a more common approach is to enable audio recordings at pre-specified questions or a random sample of questions.

Thissen and Myers (2016) identified four ways in which CARI may uncover potential falsification. First, the audio recordings may capture clicking noises but no voices, suggesting that responses to questions are being entered but no actual interview is taking place. Second, an interviewer's voice may be heard but no respondent voice is captured on the recordings. Third, a voice other than the interviewer's may be captured on recordings for multiple interviews. And fourth, there may be a high rate of missing audio recordings in cases where consent to the recordings was supposedly given. It is important to note that each of these scenarios raises suspicion but does not produce definitive evidence of falsification. For example, a laptop microphone may be obstructed, making the interviewer's and/or respondent's voice inaudible. CARI can also be used to capture partial falsification. Reviews of question entries in combination with the audio recordings may uncover instances in which an interviewer deliberately mis-enters a response to avoid a long list of subsequent questions (Edwards, Maitland, and Connor, 2017). Low rates of interviewer-level respondent consent may also indicate possible falsification (Edwards et al., 2017).

The use of GPS data is an emerging but still uncommon approach for monitoring interviewer behavior and detecting falsification. GPS coordinates can be captured in two ways: active capture by the interviewer and passive capture by software during survey administration (Keating et al., 2014; Thissen and Myers, 2016). Active capture acts more as a deterrent to falsification as it makes the interviewer aware that location data are being collected and reviewed by survey managers. Regardless of approach, the GPS coordinates are tagged (geotagging) to the household listing and/or survey responses and algorithms are used to compute distances between the actual GPS readings and the expected coordinates. Outliers are flagged for further investigation. Applications such as Google Earth can be used to provide visual comparisons between the GPS-based and expected locations. In essence, the GPS data provide verification that listers or interviewers were at a particular location at a particular time (Edwards et al., 2017; Ellis, Sikes, Sage, Eyerman, and Burke, 2011; Eng et al., 2007; Thissen and Myers, 2016).

Edwards et al. (2017) described a GPS-based system developed by Westat, an SRO, for increasing the efficiency of field operations and detecting possible falsification. Efficiency Analysis through Geospatial Location Evaluation (EAGLE) covers two approaches for integrating GPS data into field household survey operations. The first involves GPS logging[4] in which the path of travel coordinates are logged, allowing survey managers to track or trace the movement of interviewers in the field. The second approach uses a web-based mobile application (mFOS) to capture travel activity and key events, with the captured data integrated with online services for near real-time data updates. Information from both are geotagged and uploaded when interviewers submit their completed cases. Once received, the GPS point data are summarized and then processed to "determine the location of the GPS points in relation to the sampled address or person; the time when the interview data were collected compared to the GPS point at the same time; and the location of the interviewer's home" (Edwards et al., 2017, pp. 266-267). Outlier reports from EAGLE identify suspect cases that are further investigated.

Dajani and Marquette (2015) described the use of GPS data to detect what they labeled "curbstoning clusters" as part of the 2010 census address canvassing operation. A curbstoning cluster is defined as six or more housing units, excluding multi-housing unit structures, located in a square or adjacent squares of about 40 feet, which is implausible and suggestive of falsification. During the field operation, curbstoning clusters were identified at single houses, coffee shops, or spots along a road. Overall, they found that 19,500 canvassing assignments had at least one curbstoning cluster.

GPS data can also be used to deter potential falsification by alerting an interviewer if they appear to be too far from a sample unit. Geofencing, for example, is a tool in which survey managers can set virtual perimeters or geospatial boundaries around sample units, alerting interviewers if they appear to be outside those boundaries at the time of an interview (Keating et al., 2014). Alerts or warnings can be sent to interviewer cell phones or other mobile devices (Burke, 2015; Dajani and Marquette, 2015).

[4] Thissen and Myers (2016) also use the terms "geotracking" and "geocaching."

Screen captures and still images can also be used to deter and detect possible falsification (Thissen and Myers, 2016). Screen shots captured during the interview can be coupled with CARI recordings, for example, facilitating the work of coders who can compare images of the questions and keyed responses with the audio recordings. Thissen and Myers (2016) describe a survey of fishing activities conducted by RTI International in which interviewers were instructed to take photos of the actual fish and fishing environment. As they note, falsification becomes more difficult in such a survey. To further verify the location of an interviewer at the time of interview, photographs can also be geotagged (Keating et al., 2014).

An advantage of detection methods such as CARI or geotagging is that falsification can be identified and handled quickly. Interviewers can be immediately removed from data collection, pending further investigation, ensuring that falsification is minimized. In addition, some falsified cases may be reassigned and worked appropriately before data collection ends. Low cost and ease of collection allows GPS coordinates to be appended to 100 percent of the sample cases, a major advantage over methods such as recontact or reinterview that may only be applied to a small fraction (e.g., 5%) of the sample. This can also serve a deterrent role as interviewers are aware that 100 percent of their workload is being monitored in near real time as opposed to a small portion of their workload being recontacted at some point during data collection.

4.2.2 Statistical Detection Methods by Reviewing "Outputs"

The limitations of direct monitoring methods, such as RI or audio review (mentioned in the previous subsection), can be reduced by monitoring the outputs of the survey process itself (i.e., monitoring the substantive data, the paradata, and fieldwork indicators). Output monitoring potentially covers all cases without incurring incremental costs and can lead to better direct monitoring strategies (AAPOR, 2003; Hood and Bushery, 1997; Winker, 2016). While meant to cover all cases, output monitoring can also be used to focus on (1) specific items, such as gateway questions if they are known to lead to a lengthy module and, therefore, susceptible to falsification (Thissen and Myers, 2016); (2) specific interviewers, such as new interviewers (AAPOR, 2003); and (3) specific time periods, such as the start-up and concluding phases of the survey field period (AAPOR, 2003).

The key principle behind output monitoring is to relate the data in question to a reference distribution that is unknown (or difficult to know) to the falsifying agent. For example, an interviewer does not typically know the responses garnered by other interviewers and even if such access were available, it would be difficult to falsify data in a manner that closely matches all the properties of the reference distribution (especially a joint data distribution). Similarly, it is difficult to create duplicate records while keeping in mind the laws of probability by, say, avoiding duplicating rare combinations too often. Six output monitoring methods are briefly described next.

4.2.2.1 Methods Involving Descriptive Statistics

Perhaps the archetypical form of output monitoring involves the comparison of mean values or frequency distributions and analyzing deviations of a data point from a measure of central tendency. For example, Murphy, Eyerman, McCue, Hottinger, and Kennet (2005) showed that the lifetime substance-use rates from falsifier interviews were statistically lower than those obtained from valid cases. Since falsifier interviewers can be good at guessing univariate response distributions (Schraepler and Wagner, 2005), a more effective method is to compute deviations within demographic or behavioral strata that are likely to show stronger differences. Murphy et al. (2005) computed a score for each interviewer based on deviations of substance use rates within age, sex, and Hispanic-status strata. They found that the response deviations for three known falsifiers were the highest among the overall pool of 10 interviewers. Such response deviation-based methods could be applied to paradata and fieldwork productivity indicators as well. Turner, Gribble, Al-Tayyib, and Chromy (2002) found that falsifying interviewers obtained a higher proportion of eligible households and a higher proportion of completes than the "not suspect" interviewers. Murphy et al. (2016) gave a detailed list of 36 indicators that one could adapt to one's situation, such as interview lengths and keying times (falsifier interviewers tend to speed) (Birnbaum, DeRenzi, Flaxman, and Lesh, 2012; Thissen and Myers, 2016; Winker, 2016) and refusal rates for audio recording (Thissen and Myers, 2016). A related method is to set up SPC charts to monitor deviations over time (Murphy et al., 2005). For example, Bushery et al. (1999) used SPC charts with 3-σ control limits to compare weekly interviewer performance measures to a historical average. Another form of the descriptive-based method is to look at the occurrence of rare combinations within each interviewer's workload (Birnbaum et al., 2012; Faranda, 2015; Porras and English, 2004).

Descriptive-based methods have their limitations. First, if interviewer workloads are small, then one may end up with an abundance of false positives (Winker, 2016). Second, as pointed out earlier, looking at a univariate analysis might not help since interviewers may be good at faking realistic response distributions. While stratification can help, looking at too many interlocked strata may not be feasible. Third, in the absence of an interpenetrated design (Mahalanobis, 1946) (i.e., random assignment of cases to interviewers), response differences between interviewers could simply be the result of differences in geographic or respondent characteristics. Some of these limitations can be addressed by explicit model-based methods described next.

4.2.2.2 Variations in Response Distributions

In a substance-use survey experiment, Inciardi (1981) hypothesized that interviewers who were experienced with substance surveys develop a stereotype of the population and are led by this stereotype to generate data that vary less than what one would expect. To test this hypothesis, Inciardi compared valid data with fake data generated by experienced interviewers; the interviewers were not told the purpose of the exercise. Four items from the survey were then selected for analysis. The results showed that while the mean values from the fake interviews were close to those from the valid interviews, the former produced data with less variation than the latter. Blasius and Thiessen (2013) suggest that this method is particularly suitable for surveys focusing on specific groups such as immigrants or minorities since the interviewers likely referenced pre-conceived ideas or stereotypes about these groups when generating the fake data. Falsifier interviewers may also end up with less variability in their data based on the deviation-based measures described earlier (Crespi, 1945; Porras and English, 2004). Porras and English (2004) and Schäfer, Schräpler, Müller, and Wagner (2005) found the variation-based method useful in detecting falsifiers.

Not all items would be suitable for this method. For example, Inciardi (1981) notes that if one were studying patterns of sexual behavior within first-year high school populations, one could expect small standard deviations about certain variables. In practice, one would compare the variation in responses for a candidate set of items across interviewers.

Beginning with Kuriakose and Robbins (2016), the maximum percentage of questions for which each respondent matched any other respondent, referred to here as the maximum percentage match statistic, has been considered as an indicator of potential data fabrication. A maximum percentage match of 100 percent indicates that the responses are an exact match (i.e., duplicates). For a given survey, the probability distribution for the maximum percentage match may be estimated; this distribution is well described by the Gumbel distribution. For example, in a hypothetical survey of 100 questions with independent binary responses, each with a mean of 0.5, it would be highly unlikely to obtain more than an 85 percent maximum percentage match between any two respondents, whether their responses were independent or correlated. In addition, the distribution of the maximum percentage match for all respondents was well described by the Gumbel distribution.

As discussed in section 2.2.5, Kuriakose and Robbins (2016) used the results of their simulation studies to assess potential fabrication in 1,008 national surveys conducted in 154 countries over a period of 35 years. They noted that these methods work best for lengthy surveys. And, they may not work well for surveys outside the social sciences or more specifically for surveys other than public opinion surveys. They specifically noted that duplication may not be rare for customer satisfaction surveys and worker satisfaction surveys.

In summary, use of the high percentage matches may be a useful metric for assessing potential data fabrication. However, a better understanding of how this measure varies with survey characteristics is important to avoid false positives, which lead to false claims of data fabrication.

4.2.2.3 Data Reduction Methods

Two common data reduction methods are principal component analysis (PCA, and its categorical variant, CatPCA) and multiple correspondence analysis (MCA). These methods seek to reduce the dimensionality of the data by forming new dimensions that are combinations of original variables. A small number of these new dimensions typically capture most of the information in the data. Each case in the data is assigned a "score" for the new dimensions. Blasius and Thiessen (2012, 2015) used MCA to identify the presence of duplicates and unanticipated patterns in survey data. Specifically, they used the fact that respondents who get identical scores must also have identical combinations of values in the original (categorical) data. If there were a large number of initial variables, it is unlikely that a particular combination of values would show up with large frequency in the data. Consequently, the frequency of occurrence of any particular score should also be low. Analyzing three different sets of items on the German ALLBUS survey, they were able to efficiently spot respondents with identical scores who were then also found to be clustered among specific interviewers. This suggests that the interviewers might have simply pasted records from other interviews. Some of the original combinations also did not make substantive sense, which suggests that all such data were faked. Further, the researchers suggest singling out interviewers with "outlying" mean dimension scores since this suggests unnatural response patterns obtained by these interviewers.

4.2.2.4 Regression Models

Model-based methods are an improvement over the deviation-based methods described earlier. The advantage of using explicit models is that they are able to account for multiple variables simultaneously, allow for inferences when interviewers have small workloads by having them "borrow strength" from other interviewers, allow for specification of multiple levels of falsification (e.g., interviewer and supervisor), are flexible, and are capable of being transported to future survey waves (if the essential conditions remain the same).

If data from previous RIs or investigation outcomes are available, a logistic regression model can be fitted with the outcome being a binary "falsification status" variable that is predicted by a range of input variables (Li et al., 2011). This model could then be used to predict the "falsification propensity" (Murphy et al., 2016) for the current data collection period. The input variables could be demographic data, paradata (Li et al., 2011), response behavior indicators such as rounding and frequency of mid-point selection (Kemper and Menold, 2014; Menold and Kemper, 2014), and item characteristics (Menold and Kemper, 2014). In a Bayesian context, it may also be possible to incorporate prior knowledge, such as the expected distribution of valid responses from a prior survey wave, into the models.

Models can be used even in situations in which RI was not used for the survey. Sharma (2016) reported on a television viewing panel survey in India where data were collected electronically from respondents; interviewers were involved only in recruiting respondents, collecting initial demographic data, and being the point of contact for the respondent in case of technical difficulties. Since the survey data had a high commercial impact (TV channels with better viewership could charge more for advertising), interviewers were being contacted/threatened to persuade their respondents to falsify data by, say, leaving the TV tuned into a certain channel even when no one at home was watching. To detect such falsification, the survey organization used multilevel models where longitudinal data were clustered within households that were in turn clustered within interviewers. The outcome variable was the amount of viewing of a channel that was under suspicion. The models allowed for random interviewer and household intercepts and slopes, and respondent covariates were added to approximate an interpenetrated design. By examining the conditional modes of the random effects, the survey organization was able to zero in on a few interviewers and further investigation confirmed that cases within these interviewers were problematic.

Finally, Landrock (2017) shows that theory-driven complex relationships are difficult for falsifiers to reproduce. Therefore, one possible method is to fit models where the input and outcome variables are those that have theoretically known relationships. These models will have interviewer-varying slopes and will also approximate an interpenetrated design as described above. Very large or small conditional slopes would indicate those interviewers for whom the associations among variables differ from the actual association. Such results may be due to interviewer effects unrelated to falsification, and the assumption here is that the majority of the interviewers are producing good quality data. However, if differences in association remain after controlling for interviewer effects to the extent possible, other potential sources of variation, including falsification behavior, may be indicated.

4.2.2.5 Leading Digit Methods

The most well-known version of this class of analyses is based on Benford's law, an empirical principle – theoretically justified in Hill (1995) – underlying the frequency distribution of leading digits in many numerical datasets. The distribution of leading digits in such data is not uniform as one might expect, but follows:

$$Pr(\text{first significant digit} = d) = log_{10}\left(1 + \frac{1}{d}\right)$$

This is a useful result in the context of falsification since interviewers may falsify data expecting the leading digit “1” to appear about 11 percent in the data (1/9 of the time), but Benford’s law would actually indicate an estimate of 30 percent. In practice, one would extract the leading digits of data from each interviewer and compute a χ^2 statistic as follows:

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{\left(p_{d_i} - p_d^{(B)}\right)^2}{p_d^{(B)}},$$

where p_{d_i} is the proportion of leading digit d for interviewer i who has conducted n_i interviews, and $p_d^{(B)}$is the proportion of digit d under the Benford distribution. Since this statistic is dependent on the number of interviews conducted by each interviewer, Schraepler and Wagner (2005) use a relative goodness-of-fit statistic equal to $1 - \frac{\chi_i^2}{\chi_0^2}$, where χ_0^2 is the value of the distribution with the worst fit to the Benford distribution; low values, therefore, indicate a bad fit. Another version by Schräpler (2010) uses bootstrapping to remove the effect of an unequal number of interviews. However, simulations by Bredl et al. (2008) suggest that the fit to the Benford distribution is independent of sample size.

To judge the effectiveness of Benford’s law as a detection method, Schäfer et al. (2005) and Schräpler (2010) use data from the SOEP where they already knew which cases were falsified. The method based on Benford’s law was successful at detecting a large number of falsifiers. The success of this method may partly be due to the fact that the distribution can be difficult to fake (Winker, 2016).

Benford’s law, however, does not apply to all data. Scott and Fasli (2001) studied 230 datasets available on the Internet cumulating to about 500,000 numbers. They found that data from only 29 datasets conformed to Benford’s law at a 5 percent significance level. This is despite the fact that they deliberately sought data that they thought would fit the Benford distribution well. Schräpler (2010) summarizes results from several theoretical, empirical, and simulation studies such as Hill (1995), Nigrini (1999), Scott and Fasli (2001), and Mochty (2002) that suggest conditions when Benford’s law applies: the data are naturally occurring (i.e., they are not “assigned” such as school IDs), are raw in the sense of not being based on numerical summaries, consist only of positive values with a unimodal distribution, have a positive skew, and do not have a built-in maximum.

Survey data have additional issues stemming from respondents rounding their responses to the leading digit (Porras and English, 2004; Schräpler, 2010; Swanson et al., 2003), and homogeneity of respondents within interviewer assignments leading to similar responses (Schräpler, 2010).

In situations when Benford's law does not apply to the leading digit, one option is to use the first two digits (Schräpler, 2010). However, response rounding can impact the usability of the results (Bredl et al., 2008; Schräpler, 2010). Another option that assumes that most interviewers are not engaging in falsification is to replace the Benford distribution with another reference distribution. This is based on research by Scott and Fasli (2001) that showed many datasets did not conform to Benford's law but still had similar patterns of leading digits. A version of this method is to use the empirical distribution across all cases as the reference distribution (Swanson et al., 2003). Porras and English (2004) compared the leading digit distribution of data for a specific interviewer suspected of falsification to the distribution from the rest of the interviewers at different data collection stages via a chi-squared test. They found that this particular interviewer had the lowest *p*-values at each stage, indicating the effectiveness of this method.

An alternative to Benford's Law is to consider the distributions of *trailing digits*. Silver (2009) compared the distribution of trailing digits between two polling firms, Strategic Vision and Quinnipiac, in an effort to assess whether some of the Strategic Vision data may have been fabricated. The two polling firms are similar, though not identical, in their geographical coverage, length of survey instruments, and frequency of reports. For the Quinnipiac data, the trailing digits appeared to be slightly nonrandom with the smaller digits being observed slightly more often. This was likened to Benford's Law in which the smaller *leading* digits are more likely to occur. In contrast, the opposite pattern was observed in the Strategic Vision data, with the larger trailing digits occurring more often. In addition, the gap between the smallest number and largest number of occurrences of a digit was much larger. This was further explored by comparing both observed distributions to the uniform distribution. For the Quinnipiac data, the number of 2s deviated from the uniform distribution by about 2.4 standard deviations. In contrast, for the Strategic Vision data, the number of 1s departed 5.7 standard deviations from that anticipated under the uniform distribution. With these results, Silver argued that the Strategic Vision data are more nonrandom than the Quinnipiac data. Thus, he concluded that it is highly unlikely "that the distribution of the results from the Strategic Vision polls are reflective of any sort of ordinary and organic, mathematical process." The applicability of the trailing digits approach is likely to be less than that of the leading digits, especially since responders often round when providing answers to survey questions.

4.2.2.6 Cluster Analysis and Other Data Mining/Machine Learning Methods

The last couple of decades have seen an explosion in computing power and developing new methods to detect patterns in data. Some of these methods are "unsupervised" in that there is no explicit outcome variable as in a regression model. One such unsupervised method is cluster analysis that divides the data into a pre-specified number of distinct clusters based on user-defined indicators. Bredl et al. (2012) explore the idea of using the χ^2 values from the Benford distribution in conjunction with three other indicators in a cluster analysis to detect falsification. These three indicators were: a "nonresponse ratio" (the expectation was that falsifiers would have a low ratio since they tend to answer every question by pretending to be sincere), an "extreme-answers" ratio (the expectation was that falsifier interviewers will have a lower ratio than the rest), and an "others ratio," which is the proportion of relevant questions where the "other" category was marked as a response (the expectation was that falsifier interviewers record fewer "other" options since doing so would mean additional work in asking for an alternative from the respondent or falsifying even that response).

The researchers had access to a survey with 250 interviews conducted by 13 interviewers where re-interviews revealed that 4 interviewers had faked their data. They conducted a two-group cluster analysis based on the four indicators, using both hierarchical clustering and K-means clustering. They find that all methods yield a "falsifier" cluster but with varying degrees of false positives and false negatives. Further analysis showed that the "nonresponse ratio" and the "others ratio" were the best inputs to the clustering.

The advantages of using cluster analysis for this research were that there were a relatively small number of interviewers, and the falsifier interviewers engaged in fairly extreme falsification. In practice, and especially when there are many interviewers but few falsifiers, one would run into the following challenges: the clusters might not result in a neat separation; after running the cluster analysis, one would not automatically know which cluster really contains the falsifiers – more data analysis would be needed; and finally, even if one knew the "falsifier cluster," the falsifier interviewers in the cluster themselves are not identified (given that the cluster is most likely to also contain false positives). Despite these challenges, adopting cluster analysis as a screening method can be valuable if strong indicators and strong mechanisms are at play (e.g., one can imagine a cluster analysis using paradata where one cluster might contain all interviews that were done in a very short time with a very small nonresponse ratio).

A comparison of unsupervised and supervised approaches to detect falsification was conducted by Birnbaum et al. (2012). The supervised algorithms used were logistic regression[5] and random forests, and the unsupervised algorithms were local correlation integral, multinomial model algorithm, and the *s*-value algorithm. The results show that all five algorithms were able to predict falsification with at least 80 percent accuracy. This was the case even in the scenario when interviewers who were generating the fake data were told how the algorithms would work. However, the falsification rates in this research were quite large; at least one study had a falsification rate of 37 percent.

Four points need to be borne in mind with respect to output monitoring. First, as mentioned earlier, these methods cannot lead to a "confirmation" of falsification. These are meant to act as good filters so as to spot issues that can be further investigated. Second, these methods still require human interpretation and should not, therefore, be run in a fully automatic fashion (Thissen and Myers, 2016). Third, while the list of methods can seem overwhelming, not all methods will be applicable for a given situation. For example, if duplication of records is not a concern, then the data reduction methods might not be relevant. A more focused approach to detecting falsification can occur when thought is given to falsification mechanisms that are most likely to occur for the survey organization. These can inform the choice of methods.

4.2.3 Using Methods/Technologies in Combination

As described in this chapter, there are several types of methods and technologies that may be used for the detection of falsification. Not every method or technology is available or applicable to every situation. Some have significant costs but can yield a high level of confidence in the results. Others may be cheaper or quicker to implement but may miss important potential sources of falsification. To the extent possible, methods for falsification detection should be used in combination to achieve an optimal balance of detection and cost.

[5] The classification of logistic regression as an "algorithm" is common in the computer science literature.

Thissen and Myers (2016) recommend using multiple monitoring techniques in combination. They suggest prioritizing electronic or technological methods of monitoring, such as CARI and GPS, and starting monitoring as soon as data collection begins. The role of CARI, which requires some review of recordings that may be labor intensive, may best fit in field surveys where some aberration has already been detected using GPS and further evidence is required to understand the situation. For instance, if GPS records suggest that an interview was conducted at a location other than a sampled respondent's household, the CARI recording could be reviewed to confirm whether zero, one, or two voices are present. If zero or only one voice is detected, this would increase the confidence that a falsified interview occurred. Targeted use of CARI in this way may be more efficient than a purely random sampling of cases.

Li et al. (2011) also considered whether certain data elements may be used to better identify potential suspects of falsification for RI, improving on the existing design for the CPS. The authors examined interview outcome, interview mode, interview date, month-in-sample, and interviewer's experience and created simulations that suggested stratified, truncated, or PPS designs may be more effective and efficient. For instance, the PPS design increased the expected yield of falsified cases from 40 percent to 70 percent, depending on which model was used.

Murphy, Baxter, Eyerman, Cunningham, and Kennet (2004) discuss the use of multiple metrics to flag potential falsification. They found that if an interviewer's data demonstrate (1) a high response deviation score, (2) a high rare response combination score, (3) a low item-nonresponse rate, or (4) high variation of response time (two standard deviations below and above average), then there is a higher likelihood that he/she has fabricated the data.

The Latin American Public Opinion Project (LAPOP) developed a system to compile multiple sources of data for real-time analysis. This system, named the Fieldwork Algorithm for LAPOP Control over Survey (FALCON), monitors interviewers' locations at the time of interview, confirms interviewer identities, verifies respondent demographics, conducts question timing analysis, and records interviews via CARI (Robbins, forthcoming). FALCON employs an algorithm to identify outliers in each of these data types for review and investigation by the project team.

4.3 Fitting the Methods to the Survey at Hand

Winker (2016) noted that datasets known to have been subjected to at least some falsification can provide a testbed to develop and evaluate methods for detecting falsification. Although some journals require making data available as a condition of publication, this is not universal. Further, concerns of confidentiality and, in the case of Federal agencies, legal requirements may limit the availability of data, even for journals with the requirement. To address at least some of these concerns, Winker suggested the creation of a joint data research center that would guarantee data protection, both with respect to survey content and the parties associated with the data. Encouraging organizations to publish their successes and failures with detecting falsification, even if this requires anonymization of the specific survey and indexing actual data, will help other organizations learn from these experiences.

5. Impacts of Falsification on Study Results

5.1 Introduction

Data falsification negatively impacts the credibility of a research study. If the falsification rate is known to be extremely small and, thus, the effects on estimates and inferences minimal, the conclusions drawn from the analysis may continue to be questioned. If the offending data points are removed, doubts may remain as to whether all falsification incidents have been detected and corrected for. The resulting loss of credibility can well carry over to future survey waves (in case of a panel survey) or other surveys conducted by the same organization.

Apart from the ethical nature of the problem, those interested in the study lose confidence in the reported results because they intuitively realize that falsification could result in misleading inferences. The purpose of this chapter is to illustrate the validity and extent of these intuitions within a theoretical framework borrowed from the measurement error literature and supported by simulations. There exists a large literature on statistical measurement error in general (examples of book-length treatments are Fuller, 1987; Carroll, Ruppert, Stefanski, and Crainiceanu, 2006; and Yi, 2017) as well as measurement error specific to surveys (e.g., Biemer and Trewin, 1997; Bound, Brown, and Mathiowetz, 2001; and West and Blom, 2016). However, as noted by Bound et al. (2001, p. 3723), these studies largely focus on mechanisms that impact every observation in a study. Falsification is different in that it typically impacts only a subset of observations and, therefore, deserves special study. Also, in most applications, measurement errors are typically assumed to be generated at random from a zero-mean distribution. Suppose that interviewers are selected at random from the population of all interviewers, which is made up of non-falsifiers and falsifiers. Further assume that each of the falsifiers may falsify interviews at a different rate that could vary over time. Then the rate of falsification has a random component. When some falsified data are present, they may have a mean that differs from the corresponding population values, resulting in bias. Falsification is, thus, a special case of measurement error and is the focus of this chapter.

Perhaps due to the nature of the problem (e.g., unavailability of information on falsified cases, lack of true variable scores), the existing literature on the empirical findings of the impacts of falsification on study results is limited. Generally, the few papers dealing with how falsification affects analyses compare data from interviews that are known to be falsified with data from the rest of the interviews, which could also include interviews that were falsified but went undetected.

Schnell (1991) finds that a 7.2 percent falsification rate impacted regression analyses but descriptive statistics were not significantly impacted; the 220 falsified interviews in this study were "artificially" created by university students and faculty. Schraepler and Wagner (2005) find falsification rates ranging from 0.1 percent to 2.4 percent in different waves of the German Socio-Economic Panel Study (SOEP) and discuss impacts of these on descriptive and analytic statistics, including some mathematical derivations. In an analysis of social network data, Bruderl, Huyer-May, and Schmiedeberg (2013) find that 330 interviews (3.6% of respondents) belonging to "fraudulent" interviewers (a term they use for interviewers who skip or shorten the network generator question) have an average network size of only 1.6 persons as compared with the overall average of 4.3 persons.

Landrock (2017) uses a quasi-experimental design in which 78 interviewers first conducted real interviews with respondents (9 interviews on average) and were then asked to falsify data in a laboratory environment after being given a hypothetical respondent's sociodemographic information. The resulting means and proportions based on falsified interviews are not very different from the non-falsified interviews; however, the falsified data perform worse in models in generating data in line with existing social science theories. This research also used university students as interviewers who were free to choose their respondents (who were also university students). A limitation in these kinds of studies is a confounding of sampling error with measurement error; the assumption is that the distributions of non-falsified data and falsified data from the sample are equal to those that would have been observed in the population had a census been undertaken with the same rate of falsification.

5.2 Approach

In this section, theoretical findings from the measurement error literature are adapted to the specific problem of falsification; these findings are supplemented with simulations and numeric illustrations to provide more insights into how falsification can affect study results. The goal is not to give an exhaustive treatment to the subject but to underscore the fact that falsification can have an impact on the results from even simple analyses (exploratory data analysis, means, proportions, and simple linear regression) under simple designs. For the simulations and illustrations used here, the following assumptions are made:

- The study is based on a probability-based survey that uses simple random sampling without replacement;
- The sample size (n) is a small fraction of the population (N) so that the finite population correction factor can be ignored;
- Sample cases are equally and randomly assigned to interviewers so that, on average, every interviewer's workload (w) has the same composition (i.e., an interpenetrated design) (Mahalanobis, 1946); and
- Interviewers obtain a 100 percent response rate so that the number of respondents is equal to the sample size; and all measurement error is only due to falsification.

The theoretical results presented next provide a foundation for understanding the impacts of data falsification on study results, confirming common intuitions. As will be seen, there can be substantial impacts even with the admittedly simple assumptions and scenarios considered here.

In this section, the focus is on falsification due to the interviewer since this is the mechanism that is perhaps responsible for the largest share of falsification; however, the impacts on analyses are not dependent on the source of falsification as long as responses are perturbed in the manner described. The impacts of falsification on both bias and variance are studied; falsification is often considered as a biasing factor on point estimates, while its impact on variance seems to be less explored. The discussion of bias and variance is in a repeated sampling context. When discussing the bias of a sample mean, the reference is not to the difference between the population mean and the mean computed on a *single* sample; it is to the conceptual difference between the population mean and the mean of the distribution of all possible sample means (technically, the "expectation" denoted by the operator E) when samples with the same size and design are repeatedly sampled from the population. While bias is a systematic error that shows up in repeated samples, variance is random error describing the spread of the data about the mean. Only minimal mathematical detail is provided in the following discussion; readers can consult the references for more detail.

5.3 Impact of Falsification on Exploratory Data Analysis (EDA)

Most statistical analyses begin with EDA (Tukey, 1977) marked by "a willingness to look for those things that we believe are not there, as well as those we believe to be there" (Jones, 1986, p. 806).

Falsification can mislead on both counts. In a hypothetical example, interest lies in the number of hours the television is turned on daily in a household ($n = 1{,}000$) and whether it differs for two demographic groups based on the number of members in the household. The values for the number of hours the television is turned on daily were generated from two normal distributions with means of 4 hours and 10 hours and standard deviations of 2 hours each (that is, $N(4{,}2)$ and $N(10{,}2)$. Based on the true sample values, the variable's bimodal distribution is evident (left panel of Figure 5-1).

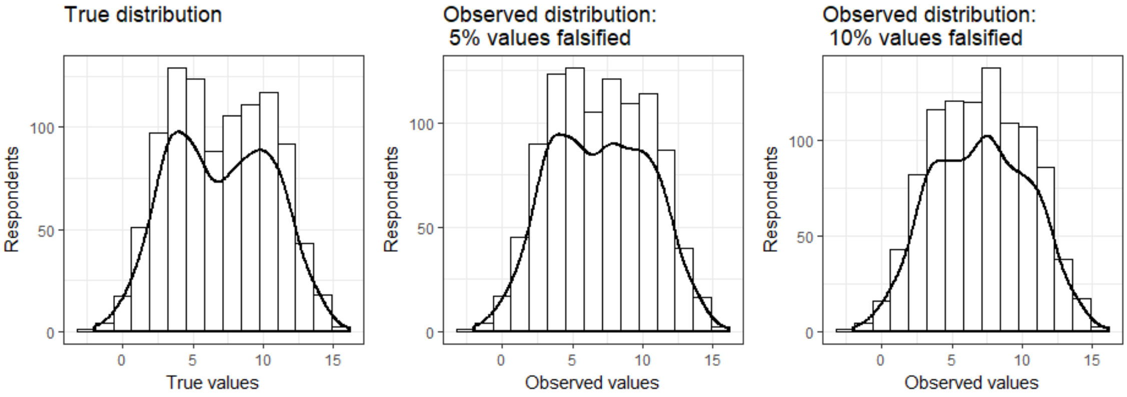

Figure 5-1. Comparison of the distribution of true values for a variable (left panel) with distributions where 5 percent and 10 percent values have been falsified (center and right panels). Falsified values mask the bimodal nature of the true distribution.

Now, take two scenarios in which 5 percent and 10 percent of the values in the study were falsified. These values were generated from a $N(7{,}1)$ distribution corresponding to a falsification mechanism where interviewers are trying to be "safe" by reporting values near the center of the overall variable distribution. The plots based on the falsified data are shown in the center panel (5% falsification) and right panel (10% falsification) of Figure 5-1. Even with a 5 percent falsification rate, the bimodal nature of the distribution is suppressed. With a 10 percent falsification rate (100 values falsified), the distribution is completely distorted. This example illustrates how falsification can mask features of the distribution especially when it occurs in the same region of the data that delineate these features. If an objective of the study was identifying sociodemographic factors associated with the number of hours a household's TV is on daily, an important factor (family size) would be missed. And, if the objective was prediction of the hours based on various sociodemographic factors, the analyst could miss including "number of members at home" as a predictor in the model if she went by the EDA based on falsified data.

Depending on the falsification mechanism, a reverse situation can also occur where "ghost features" are introduced for example, an analyst plots the relationship between two variables x_1 and y_1 for a domain with $n = 200$ ["true" data were generated from $\boldsymbol{x}_1 \sim \mathrm{N}(3,1)$ and $\boldsymbol{y}_1 = 3\boldsymbol{x}_1 + \boldsymbol{\epsilon}; \boldsymbol{\epsilon} \sim \mathrm{N}(0,1)$]. The true linear relationship between $\boldsymbol{x}_1$ and $\boldsymbol{y}_1$ is evident in the plot of the data (left panel of Figure 5-2). In the falsified version of the data, 27 x_1 values that were less than 2, were replaced by values from a uniform distribution between 1.9 and 3. The true relationship is now replaced by a feature that appears more curvilinear (right panel of Figure 5-2).

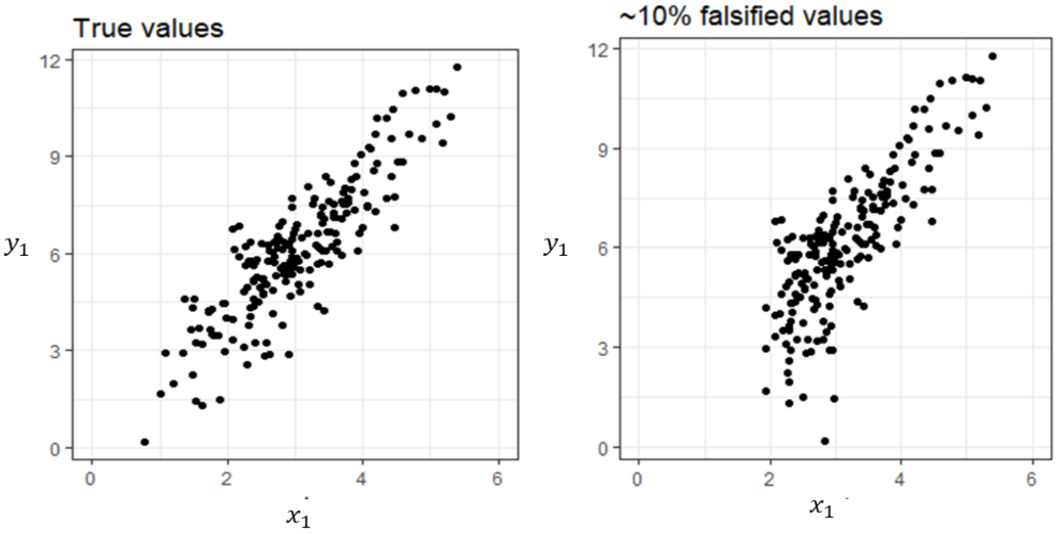

Figure 5-2. Scatterplots of the relationships between x_1 and y_1. The true relationship is shown in the left panel and that from the falsified data in the right panel.

In situations where falsification is stark, falsified data points can appear as clusters plainly visible in a plot. But, an analyst would not know whether these values are erroneous or encode valuable information.

5.4 Impact of Falsification on the Sample Mean for a Continuous Variable

Intuitively, data falsification leads to bias in the estimate of the population mean of a continuous random variable **if** the mean of the distribution of falsified values differs from the population mean of the true values. In this section, insights into the quantitative impacts of falsification are provided using a simple example.

Let y_{ij} be the value of a continuous survey variable recorded for a question administered by interviewer $i = 1, \dots, I$ to respondent $j = 1, \dots, w$ ($w = n/I$, the interviewer workload). A "falsification model" – a measurement error model that suggests the mechanism by which falsification values occur – helps illustrate how falsification affects study results. One example is an additive model as follows:

$$\textit{falsified record value} = \textit{true recorded value} + \textit{discrepancy}$$

$$y_{ij}^{(F)} = y_{ij}^{(T)} + d_i \tag{1}$$

The "F" and "T" superscripts on y_{ij} denote "Falsified value" and "True value," respectively. Given the focus here, the subscript on the discrepancy d_i is only in terms of the interviewer. Following O'Muircheartaigh (1977, p. 222), this can be interpreted as the average falsifying effect of an interviewer on their workload. The discrepancies are assumed to come from a distribution centered on $B^{(F)}$ (i.e., bias due to falsification) with a variance σ_d^2, and to be uncorrelated with $y_{ij}^{(T)}$. Now $\bar{y}^{(T)}$ is in unbiased estimator of $\bar{Y}$:

$$E\left(\bar{y}^{(T)}\right) = \bar{Y} \tag{2}$$

However, when falsification is present, the analyst actually calculates $\bar{y}^F$, a biased estimate of $\bar{Y}$:

$$E(\bar{y}^F) = \bar{Y} + B^{(F)} \tag{3}$$

The only situation where $\bar{y}^F$ will be unbiased is when $B^{(F)}$ equals zero. This will happen only when interviewers falsify cases in a way that, on average (in a repeated sampling context), the discrepancies cancel each other out – a situation unlikely to occur in practice.

The magnitude of $B^{(F)}$ depends on the falsification rate and the magnitude of the falsified values. Take the case of a hypothetical survey that contains a question on annual income. Previous (unfalsified) waves of this survey have values coming from a N($50K, $10K) distribution ("K" representing "thousands"). Now, suppose the current wave of the survey is subject to falsification. To get a sense of the impact of falsification, we varied the falsification rates (1%, 5%, 10%, and

20%) for each of three assumed falsification distributions – N($60K, $10K), N($70K, $10K), and N($80K, $10K). Note that the falsification distributions are different from the true value distributions (centered on different means); this is the situation likely to occur in practice. The sample means for these 13 data versions (one true data vector and 12 falsified versions) are computed. The above process was repeated 1,000 times. For each of the 13 versions, the mean of all 1,000 sample means is plotted in Figure 5-3 along with the associated 95 percent confidence interval (ignoring any possible clustering). The bias can be judged by the vertical distance of the dots from the true value. It depends on the falsification rate and the falsification magnitude, and not on whether falsified cases were concentrated within a few interviewers or spread across interviewers.

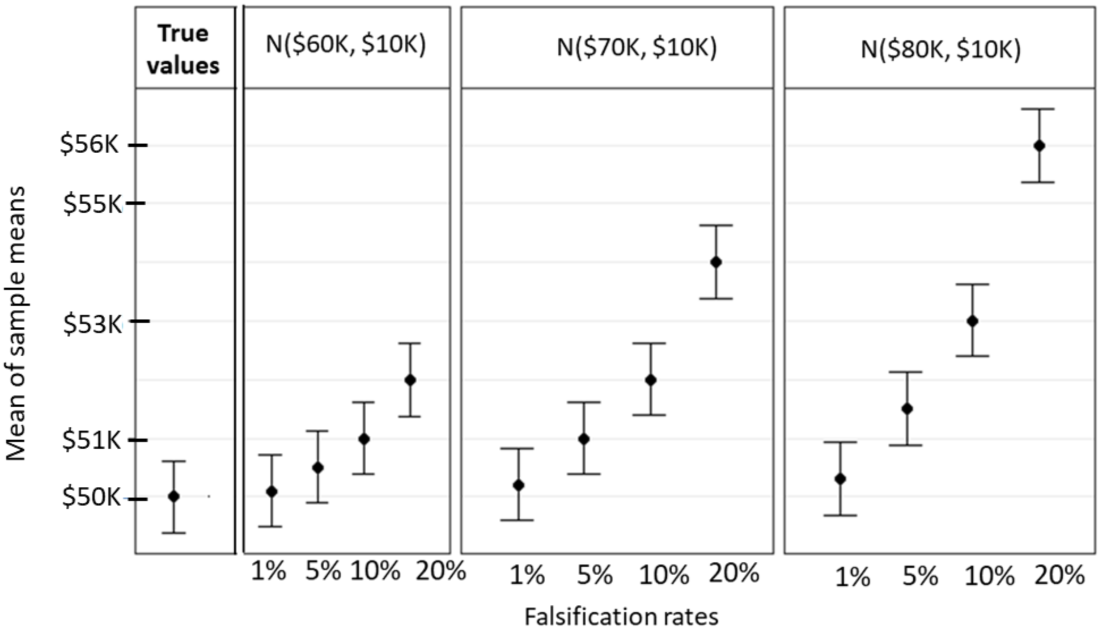

Figure 5-3. **Impact of various falsification rates and falsification magnitudes on the sample mean. The dots are means computed across the 1,000 simulated datasets for the true values and the 12 combinations of falsification rates and falsification magnitudes. The vertical bars for each dot are the 95 percent confidence intervals reflecting sampling error.**

The means are fairly robust to a 1 percent falsification rate across the three imposed falsification distributions (see Figure 5-3). However, as the falsification magnitudes increase, the difference in means for various falsification rates becomes larger. If the current wave is subject to a falsification rate of 20 percent and falsified values come from a N($80K, $10K) distribution (the fourth panel in Figure 5-3), the mean income is $56,000. This represents a misleading 12 percent increase in average income for that demographic (the true average being $50K), which could result in misdirected policy decisions.

If the population parameter of interest is a total, the biases of its estimate (y) would multiply $Bias(\bar{y})$ by the population count:

$$Bias(y^F) = Bias\left(N \cdot \bar{y}^{(F)}\right) = N \cdot Bias\left(\bar{y}^{(F)}\right) \tag{4}$$

5.5 Impact of Falsification on the Precision of the Sample Mean for a Continuous Variable

The precision of a statistic is often measured by its variance (Kish, 1965, p. 510). Under the falsification model described earlier, the variance of a sample mean is (Sukhatme and Seth, 1952; Hansen, Hurwitz, and Bershad 1961; Groves, 1989, p. 312-313; Biemer and Trewin, 1997):

$$\textit{Variance of the mean of falsified values} =$$

$$\textit{Variance of the mean of true values} + \textit{additional variance due to falsification}$$

$$Var(\bar{y}^F) = \frac{\sigma^2}{n} + \frac{\sigma_d^2}{I} \tag{5}$$

where, adapting the notation of Biemer and Trewin (1997), $Var(\cdot)$ represents the variance and $var(\cdot)$ its estimate.

A crucial point is that even if $B^{(F)}$ were to be zero, that is, the sample mean was an unbiased estimator of the population mean, the standard error of the mean when some data are falsified is larger than the standard error of the true values $(\sigma/\sqrt{n}\,)$, resulting in wider confidence intervals, which impact inferences.

From Biemer and Trewin (1997), equation 5 can also be expressed as:

$$Var(\bar{y}^F) = \frac{\sigma^2}{n}.\frac{1}{\lambda}.\left[1 + (w-1)\rho^{(F)}\right] \tag{6}$$

where the "reliability ratio" $\lambda = \frac{\sigma^2}{\sigma^2+\sigma_d^2} \leq 1$ is the ratio of variance of the true values to the total variance, w is the interviewer workload (n/I), and $\rho^{(F)} = \frac{\sigma_d^2}{\sigma^2+\sigma_d^2}$ is the correlation among values within an interviewer's workload. As the the reliability ratio decreases, the variance of the sample mean increases. When an interviewer engages in falsification, reported values may resemble each other more than what would be usual. This phenomenon is captured by $\rho^{(F)}$, which is exactly analogous to the intra-interviewer correlation coefficient (ρ_{int}) in the interviewer effect literature (Kish, 1962). The only difference is that here the focus is on one aspect of interviewer behavior, that is, falsification, hence the superscript "F" in $\rho^{(F)}$. Equation 6 implicitly assumes a nondifferential mechanism (i.e., all interviewers are prone to falsify). This might seem like a strong assumption given that many surveys might have only a small number of falsifiers. However, this does not necessarily invalidate the application here – one can still interpret $\rho^{(F)}$ as an average effect across interviewers with the non-falsifier interviewers having a very small but positive propensity to falsify. The term $1 + (w-1)\rho^{(F)}$ is a measure by which variances get inflated due to falsification (the interviewer "falsification effect"). Equation 6 shows that even a small $\rho^{(F)}$ value can inflate variance since it is multiplied by w; as an illustration, if $\rho^{(F)} = 0.01$ and $w = 100$, the variance is doubled as compared to an estimate without the falsification.

To get a better intuition on $\rho^{(F)}$ and its impact on the precision of estimates, a simulation was conducted with the following design:

a. $y_{ij}^{(T)} \sim \text{N}(5,1)$ with $n = 1{,}000$

b. Number of interviewers: 10 and 20

c. Equal interviewer workloads: 50 for 10 interviewers and 100 for 20 interviewers

d. Falsifier interviewers randomly selected from all interviewers.

e. Number of falsifying interviewers: 2 and 5.

f. Overall falsification rates: 5 percent and 10 percent.

g. Falsified responses: $y_{ij}^{(F)} \sim \text{N}(7,1)$

h. True values for each falsified interview were replaced by falsified values; falsified values were equally and randomly assigned to falsifier interviewers.

Thus, the combinations of number of interviewers, number of falsifying interviewers, and falsification rate gives eight scenarios (see Table 5-1). For each of the scenarios, 1,000 simulations were conducted:

Table 5-1. Simulation scenarios

<table>
<tr><td colspan="6">Falsification rate = 5%</td></tr>
<tr><td>Scenario</td><td>Interviewers</td><td>Workload</td><td>No. of falsifier interviewers</td><td>Falsified cases (per falsifying interviewer</td><td>% Falsified cases (within falsifier workload)</td></tr>
<tr><td>1</td><td>10</td><td>100</td><td>2</td><td>25</td><td>25%</td></tr>
<tr><td>2</td><td>10</td><td>100</td><td>5</td><td>10</td><td>10%</td></tr>
<tr><td>3</td><td>20</td><td>50</td><td>2</td><td>25</td><td>50%</td></tr>
<tr><td>4</td><td>20</td><td>50</td><td>5</td><td>10</td><td>20%</td></tr>
<tr><td colspan="6">Falsification rate = 10%</td></tr>
<tr><td>5</td><td>10</td><td>100</td><td>2</td><td>50</td><td>50%</td></tr>
<tr><td>6</td><td>10</td><td>100</td><td>5</td><td>20</td><td>20%</td></tr>
<tr><td>7</td><td>20</td><td>50</td><td>2</td><td>50</td><td>100%</td></tr>
<tr><td>8</td><td>20</td><td>50</td><td>5</td><td>20</td><td>40%</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

As expected, the observed values of $\rho^{(F)}$ are small (see Figure 5-4). The exceptions are the scenarios involving 2 falsifiers among 20 interviewers (solid blue line). The interpretation for this is perhaps clearer when using the analysis of variance framework (Kish, 1962; Biemer and Lyberg, 2003, p. 163):

$$\rho^{(F)} = \frac{\textit{between-interviewer variance}}{\textit{between-interviewer variance} + \textit{within-interviewer variance}} \qquad (7)$$

In the 20 interviewers-2 falsifier scenario, 50 percent of the workload of each falsifier interviewer comprises false cases when the falsification rate is 5 percent. This, of course, increases to 100 percent of the workload when the falsification rate is 10 percent. This can significantly increase the mean within these interviewers (the false values come from a distribution with a higher mean than the true values) leading to a larger between-interviewer variance. With 10 interviewers, the larger workload is better able to absorb the effect of these false values, that is, a lower correlation and therefore a lower $\rho^{(F)}$.

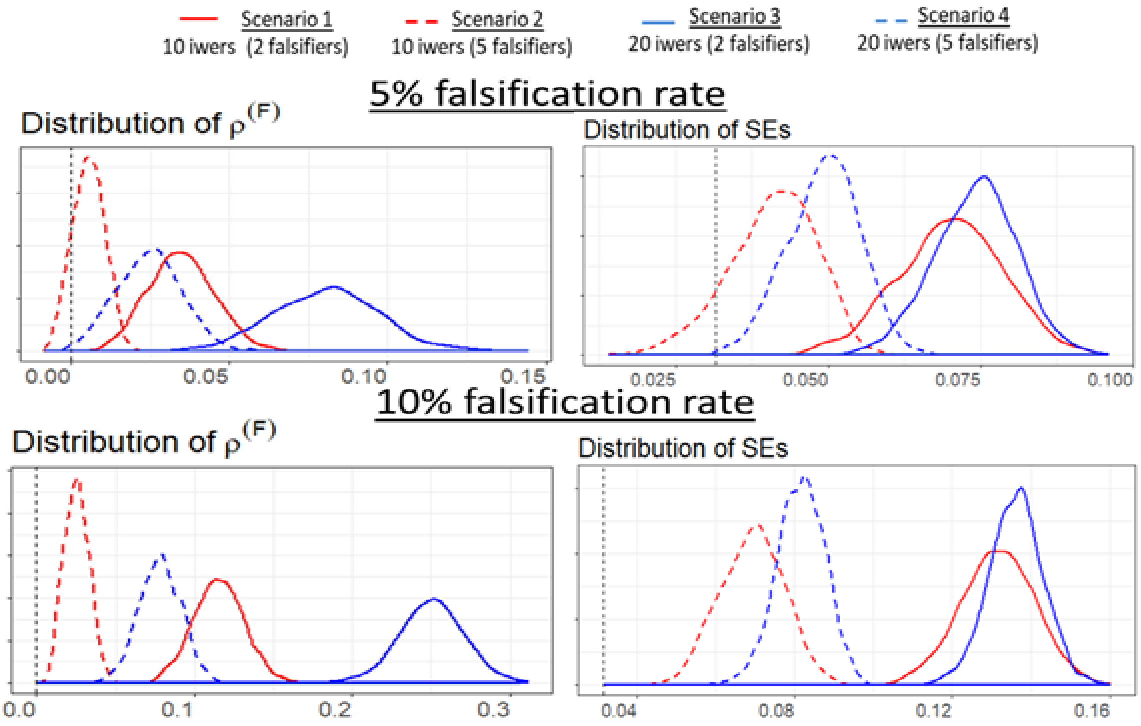

Figure 5-4. **Distribution of $\rho^{(F)}$ and standard errors for varying number of interviewers, number of falsifier interviewers, within two different rates of falsification. The dotted vertical line in each panel indicates the true values.**

This does not mean that fewer interviewers are better; a lower $\rho^{(F)}$ for the 10 interviewer cases gets multiplied by a larger w when computing the variance of $\bar{y}^F$ (equation 6). This is the reason for the large overlap in the standard error (SE) distributions (right panels of Figure 5-4). Small $\rho^{(F)}$ values cause a large inflation in standard errors relative to the true SE (compare right and left panes of Figure 5-4). For example, $\rho^{(F)}$ of 0.025 in scenario four with a 5 percent falsification rate (blue dotted line) increases the SE by 65 percent. Confidence intervals based on these standard errors will be wider than those based on true values leading to a loss of power in detecting effects.

In Figure 5-4, the curves associated with five falsifiers (dotted curves) are centered on lower values than the curves associated with two falsifiers (solid curves) seemingly suggesting that it is better to have more falsifiers. This points to an issue with the estimation of such effects; the falsification gets "normalized" across interviewers when many of them engage in such behavior leading to a lower between-interviewer component. It is important to note that in these scenarios, the overall falsification rate is assumed to be fixed – an assumption that would not necessarily hold with a given pool of potential interviewers. Also, recall that this discussion pertains to the estimation of standard errors and not bias; the bias for any scenario from such falsification will still be the same as explained in Section 5.4. Finally, with a 10 percent falsification rate, the $\rho^{(F)}$ have increased and the differences between the 2 falsifier and 5 falsifier interviewers are starker than the 5 percent falsification rate scenarios (see the plots for the two different falsification rates in Figure 5-4).

Over and above these issues, analysts computing variance estimates using the regular formula actually tend to see a smaller value of the standard error since $\rho^{(F)}$, which is generally positive, tends to introduce an underestimate of the true variance (Hansen et al., 1961; Biemer and Trewin, 1997):

$$\begin{aligned} Bias(variance\ estimate) &= E[var(\bar{y})] - Var(\bar{y}) \\ &= -\frac{(\sigma^2 + \sigma_d^2)}{n-1}\Big((w-1)\cdot\rho^{(F)}\Big) \end{aligned} \tag{8}$$

5.6 Impact of Falsification on a Sample Proportion and Its Precision

Given a binary response, errors due to falsification assume two discrete possibilities: a "false positive," which would occur when the true response was negative (or "No," 0, etc.) but the value was falsified as positive (or "Yes," 1, etc.); and a "false negative," which would occur when the true response was positive but the value was falsified as negative. The discrepancies now come from a bivariate distribution centered on (ϕ, θ), the probabilities of false positives and false negatives. The bias for the population proportion P for a variable (and $Q = 1 - P$) is given by (Cochran, 1968; U.S. Bureau of the Census, 1985; Biemer and Trewin, 1997):

$$Bias(sample\ proportion) = -P\theta + Q\phi \tag{9}$$

Analogous to the continuous variable case:

a. Unless the numbers of false positives and false negatives cancel each other out, the result will be a biased estimate of the population proportion.

b. The variance of the sample proportion based on falsified data is inflated by $1 + (w-1)\rho^{(F)}$. However, the formula for $\rho^{(F)}$is different (Biemer and Trewin, 1997, p. 616).

5.7 Impact of Falsification on the Regression Slope and Its Precision for Simple Linear Regression

Consider the following simple linear regression model:

$$y_{ij}^{(T)} = \beta_0 + \beta_1^{(T)} x_{ij}^{(T)} + \epsilon_{ij}, \epsilon_{ij} \sim \mathrm{N}(0, \sigma_e^2) \tag{10}$$

Three situations in which falsification can occur are considered: only the independent variable is falsified, only the dependent variable is falsified, and both the independent and dependent variables are falsified.

a. When only the independent variable is falsified.

When y_{ij} are measured and reported correctly but x_{ij} are falsified, an analyst will actually be fitting:

$$y_{ij}^{(T)} = \beta_0^{(F)} + \beta_1^{(F)} x_{ij}^{(F)} + \epsilon'_{ij}, with\ \epsilon'_{ij} \sim \mathrm{N}\left(0, \sigma_{e'}^2\right) \tag{11}$$

Substituting the falsification error model in equation 1 in terms of $\boldsymbol{X}$ gives

$$y_{ij}^{(T)} = \beta_0^{(F)} + \beta_1^{(F)} x_{ij}^{(T)} + \left(\epsilon'_{ij} + \beta_1^{(F)} d_{i,x}\right) \tag{12}$$

Here, the regression coefficient $\beta_1^{(F)} = \lambda \beta_1^{(T)}$ (Fuller, 1987; Carroll et al., 2006) where λ is the same reliability ratio as in equation (6). This means that the regression slope coefficient is attenuated (i.e., biased towards zero).

Looking at this in terms of the true regression line, suppose the falsification error model (equation 1) is substituted in equation 10, that is,

$$y_{ij}^{(T)} = \beta_0 + \beta_1^T x_{ij}^{(F)} + \left(\epsilon_{ij} - \beta_1^{(T)} d_{i,x}\right) \tag{13}$$

Then, assuming that $d_{i,x}$ are not correlated with ϵ_{ij}, more variation is observed around the true regression line since:

$$Var\left(\epsilon_{ij} - \beta_1^{(T)} d_i\right) = \sigma_e^2 + \left(\beta_1^{(T)}\right)^2 \sigma_d^2 > \sigma_e^2 \tag{14}$$

A simulation was conducted to visualize these results. We generate a dataset of $n = 1{,}000$ with:

$$x_{ij}^{(T)} \sim \mathrm{N}(5,1);\, y_{ij}^{(T)} = x_{ij}^{(T)} + \epsilon_{ij};\, \epsilon_{ij} \sim \mathrm{N}(0,1) \tag{15}$$

Two falsification rates (5% and 10%) and two different falsification value distributions—$\mathrm{N}(5,2)$ and $\mathrm{N}(7,1)$ are considered, resulting in four scenarios. The first falsification value distribution differs from the true value distribution in the standard deviation and the second falsification distribution differs from the true value distribution in the mean. The fitted models are shown in Figure 5-5.

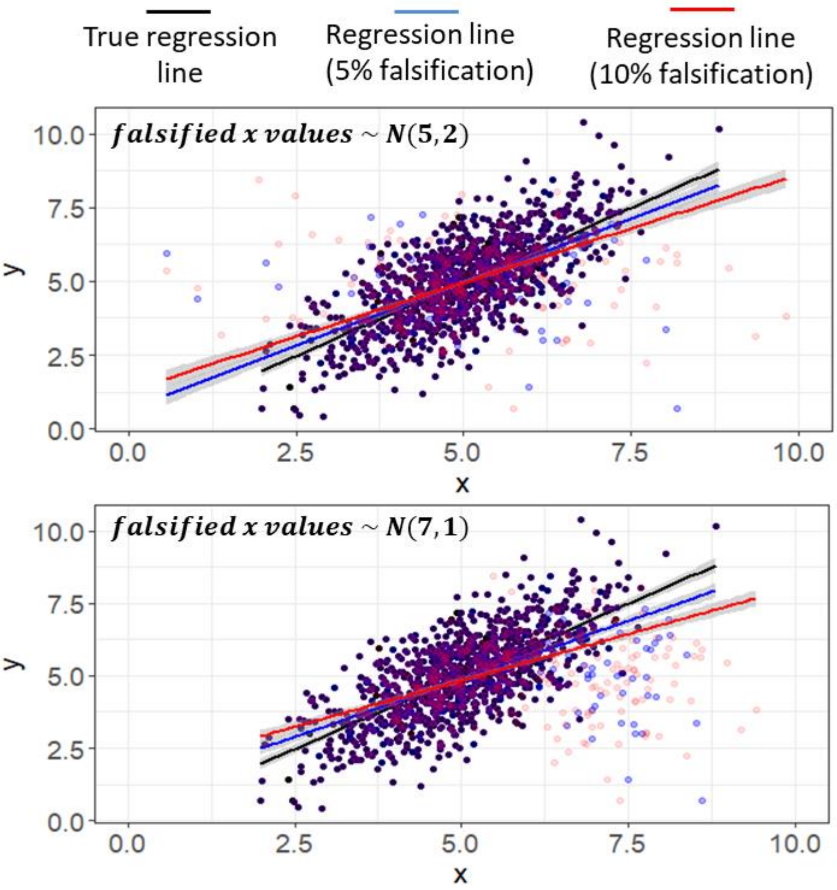

Figure 5-5. Fitted models for two different falsification value distributions. Each of the two plots has three regression lines: the true regression line (black line through the black points), the regression line with a 5 percent falsification (blue line through the blue points) and regression line with a 10 percent falsification (red line through the red points).

As displayed in Figure 5-5, falsification has made the relationship between the variables weaker. In the case of a multiple linear regression, even if only one independent variable were subject to falsification, estimates of the other regression slope coefficients would also be attenuated, the magnitude of attenuation depending on the partial correlation of the falsification error with the other variables (Bound et al, 2001, p. 3,713). Schraepler and Wagner (2005) report a comparison of regression analyses with and without falsified data where the log of gross income was regressed against age, square of age, gender, duration of training in years, and employment status. With a falsification rate of just 2.1 percent, they find the coefficient of duration of training was 40 percent lower in the full data set as compared to the coefficient in the non-falsified data, showing that the attenuation effect due to falsification can have major implications.

A higher falsification rate has resulted in more attenuation making the estimated relationship seem weaker than it is, leading to less power to detect effects (see Figure 5-5). The difference in the slopes is more pronounced when the falsified values are generated from a $\mathrm{N}(7,1)$ distribution compared to a case when they are drawn from a $\mathrm{N}(5,2)$ distribution due to a smaller λ for the former. For the $\mathrm{N}(7,1)$ falsifying distribution; the cluster of falsified points has shifted horizontally away from the cloud of points (the y values are the same) and are drawing the regression lines closer to them (lower panel in Figure 5-5).

The larger spread of the data due to falsification may be anticipated to result in a larger standard error for the slope coefficient. However, this intuition only holds if the true model error variance (σ_ϵ^2) is large or the variance of the falsification discrepancies (σ_d^2) is large or ${\beta_1^{(T)}}^2$ is small (Buzas, Stefanski, and Tosteson 2005; Carroll et al., 2006, p. 44). Therefore, with falsification, the estimate of the relationship is biased, has more variance, and has the illusion of more precision (unless the conditions just stated are met).

In illustrating the above case, any clustering of falsification by the interviewer was not considered. This does not make a difference since the design effect for the slope is dependent on clustering in both the independent and dependent variables (Neuhaus and Segal, 1993), expressed in our case as $1+(w-1)\rho_y^{(F)}\rho_x^{(F)}$, since only falsification in the independent variable, $\rho_y^{(F)}=0$ is considered.

It must be emphasized that all the results given in this section are dependent on the falsification model. For example, if the falsification model is that in equation 16, the expression for $\beta^{(F)}$ is more complex (equation 17):

$$x_{ij}^{(F)} = \gamma_0 + \gamma_1 x_{ij}^{(T)} + d_i \tag{16}$$

$$\beta^{(F)} = \frac{\beta^{(T)}\gamma_1\sigma_{x_T}^2 + \rho_{\epsilon d}\sqrt{\sigma_\epsilon^2\sigma_d^2}}{\gamma_1^2\sigma_{x_T}^2 + \sigma_d^2} \tag{17}$$

Here $\rho_{\epsilon d}$ is the correlation between the model errors and the discrepancies.

b. When only the dependent variable is falsified.

Combining equation 10 and the falsification model of equation 1 (in terms of **Y**), we get:

$$y_{ij}^{(F)} = \beta_0 + \beta_1^{(T)} x_{ij}^{(T)} + \left(\epsilon_{ij} + d_i\right) \tag{18}$$

This shows that if only the dependent variable is falsified and assuming no nonresponse error, the slope remains unchanged. However, the variance is increased due to the d_i. Recent research (Fischer, West, Elliott, and Kreuter, 2018) that looks at the interaction between measurement error and nonresponse error shows that when response propensity depends on the dependent variable, bias in the estimated slope parameter is introduced.

c. When values of both the independent and dependent variables are falsified.

When both the independent and dependent variables are falsified, there is no certainty about the behavior of the regression coefficients. Depending on how the discrepancies in the independent variable and dependent variable are correlated, the regression coefficients can be overestimated or underestimated (Chai, 1971). However, given the product of the $\rho^{(F)}$s in the falsification effect $\left[1 + (w-1)\rho_y^{(F)}\rho_x^{(F)}\right]$, the effect due to falsification will be less than that for statistics like the means or proportions (Hosmer and Lemeshow, 2000, p. 220).

5.8 Conclusion

With even the simple designs and assumptions presented here, one can see how falsification can have an impact on a study's results, negatively affecting its integrity. In fact, Sharma and Elliott (2019; a summary is in Chapter 4) reports on a survey where interviewers were suspected of being coerced or incentivized to falsify in order to directly impact the analysis of survey data. For real-world surveys where falsification behavior has complex interactions with other survey conditions, survey variables and errors correlate in various ways, and the falsification model is not straightforward. It is difficult to predict the impacts since "even in linear regression with multiple covariates, the effects of measurement error are complex and not easily described" (Carroll et al., 2006, p. 63), and “a general consensus is to conduct a case-by-case examination in order to reach a valid statistical analysis for error-contaminated data” (Yi, 2017, p. IX).

In some cases, careful analysis can contribute to negating the effects of falsification. As an example, suppose an item in a survey is prone to falsification in the form of interviewers skipping items by marking them as nonresponse. Simply undertaking a complete case analysis discards these cases, potentially resulting in biased estimates. Good imputation methods can ameliorate the situation.

Clearly the best solution is to prevent falsification and to detect and remove falsification should it occur. These steps protect the integrity of the research by ensuring that the results are unbiased, uncertainty is properly quantified, and inferences are accurate.

6. Existing Organizational Guidelines and Policies

6.1 Introduction

This chapter reviews how major professional associations address data falsification and fabrication in their codes of ethics or professional conduct and provides a snapshot of the current landscape of existing organization guidelines and policies regarding data falsification and fabrication. The codes selected for review were identified by the task force as those most relevant to survey and public opinion research but is not an exhaustive list. Additionally, the discussion of policies, practices, and guidelines used by SROs is not representative of the entirety of our profession. As will be apparent as the methodology is described, this review provides a description of the policies, practices, and guidelines of a purposive subset of SROs as a way to inform the field of approaches that are being used and/or have been found to be helpful in preventing and detecting falsification.

6.2 Review of Existing Professional Codes of Ethics and Practices

The task force reviewed the professional codes of ethics and conduct of several professional associations that are likely to have many members engaged in survey and public opinion research. Specifically, we reviewed the codes of the AAPOR, ASA, American Sociological Association, World Association for Public Opinion Research, CASRO, ICC/ESOMAR, Marketing Research Association, American Psychological Association, American Political Science Association, National Council on Public Polls, Population Association of America, American Anthropological Association, International Statistical Institute, and the Market Research Society. Additionally, some associations and professions do not have their own code but either have adopted another (often ICC/ESOMAR) or suggest several other codes as guidelines of ethical behavior.

The codes were reviewed to identify what they say about falsification regarding their members' responsibility to their clients, the public, and their profession in general. We looked for specific mentions and more general references regarding falsification. Generally, most codes had some language which, while not specifically mentioning falsification, can be interpreted to condemn falsification. Examples include:

- Researchers must not make false or otherwise misleading statements about their skills, experience, or activities (ICC/ESOMAR International Code (Article 9e); and

- If serious distortions of research are discovered, this will be publicly disclosed (World Association for Public Opinion Research [WAPOR] Code of Ethics (Section II.B.(c))).

Some of the codes reviewed are not prescriptive but suggest how members should act and include no enforcement mechanism or consequences for unethical behavior. Five of the codes reviewed specifically mention a pledge or duty to not engage in data fabrication or falsification:

- American Anthropological Association – "…researchers…should not knowingly misrepresent (i.e., fabricate evidence, falsify, plagiarize) or attempt to prevent the reporting of misconduct…"

- AAPOR – "We will not engage in data fabrication or falsification."

- American Psychological Association – "Psychologists do not fabricate data."

- American Sociological Association – "Sociologists do not fabricate data or falsify results in their publications or presentations….and do not omit relevant data."

- Marketing Research Association – "Never falsify or omit valid data at any phase of a research study or project."

Additionally, the ASA, while describing unethical behavior in its code, provides a footnote stating that "Misconduct is not limited to instances of plagiarism and data fabrication or falsification."

Many of the publicly available codes the task force reviewed are over 10 years old, some 20 years old. It is certainly possible that future revisions and updates to these may address falsification more explicitly than they do now. A report by the National Academies of Sciences, Engineering, and Medicine (NAS 2009) suggests the digital age has made it easier to manipulate data and this may tempt researchers to do so. Thus, "digital technologies are having such a dramatic effect on research practices that some professional standards affecting the integrity of research data either have not yet been established or are in flux" (NAS, 2009; p.56). Certainly, these technological advances may spur associations to address this issue when codes are updated or revised; as such, our review may become out of date as organizations revise or update codes to reflect new digital technologies.

A final consideration when thinking about current codes of conduct and their treatment of falsification is that while most discuss ethical behavior on the part of researchers, and some specifically cite a pledge not to falsify, none appear to require the researcher to take steps to ensure data integrity is protected from misconduct by others on the research team. While it may be difficult for researchers to police the behavior of others, they can try to create an environment where falsification is discouraged and pledge to publicly report it when it is discovered. Perhaps this, too, will be addressed in future code revisions and updates.

6.3 Description of Process Used to Obtain Information from Survey Research Organizations

The task force conducted a survey of select SROs to ask about policies and practices with respect to the following:

- Pre-data collection activities, such as new interviewer and human subjects' protection training, signed data integrity affidavits, interviewer and others' awareness of monitoring and verification practices;
- Real-time data collection monitoring, dashboards, and paradata review; and
- Post-data collection activities, such as statistical modeling tools.

The set of practices covered by the survey was drawn from the literature on methods, as reviewed in Chapters 3 and 4 of this report.

The survey was administered in the form of a checklist, with a series of yes/no questions asking whether the SRO engaged in the activity. While the focus of this report is on interviewer behavior, falsification can occur by staff in various roles at all stages of the survey design, data collection, and data preparation and analysis process. Thus, the questions targeted particular stages and particular staff roles. The SROs were also asked to share any relevant written guidelines or technical documents developed by the organization. Additionally, SROs were asked to provide any evidence (documented or anecdotal) of the effectiveness of their policies and guidelines. The survey request letter and the questionnaire are provided in Appendix A and Appendix B, respectively.

Because there is no comprehensive list of SROs, the task force considered what organizations should be contacted for this effort and began with the list of organizations that are members of

AAPOR's Transparency Initiative (TI; see https://www.aapor.org/Standards-Ethics/Transparency-Initiative/Latest-News.aspx). Each TI member has already signed on as being "willing to publicly disclose its basic research methods and make them available for public inspection." While these organizations may be more disposed to higher standards as a result of their willingness to join the TI, the task force believes they would provide a good snapshot of what is being done regarding falsification in our industry. Furthermore, as a result of the TI membership application process, a contact at each TI member organization has been identified, making it feasible to administer the survey request.

In addition to the TI members, a few other large SROs—major players in survey research who are not TI members—were included in the survey effort. The set of organizations that completed the task force's survey is not representative of SROs in general. Additionally, some SROs contacted for the survey did not respond to the task force's request for information within the timeframe allotted for this effort. There was never an intention for the results of this SRO survey to be generalizable. Rather, the task force aimed to describe some of the more commonly used policies, practices, and guidelines of select SROs.

The SRO survey was put in the field on April 23, 2018, with reminder follow-ups on May 7th and June 8th. Invitations were sent via email to the TI contacts (as well as contacts identified by task force members for the SROs that were added to the list) to complete the survey on the web. A total of 78 SROs were invited to complete the survey, and 14[6] completed it. Some of the invitees said that they did not collect their own data but that another organization on the list was contracted to do so. Table 6-1 provides counts of the numbers of SROs responding to the task force request, by various characteristics of the SRO. More on the survey methodology and the survey results are available in Appendix C.

[6] The following organizations completed the survey: Abt Associates Data Science, Surveys, and Enabling Technologies Division; Center for the Study of Los Angeles at Loyola Marymount University; Institute for Policy and Opinion Research at Roanoke College; Mathematica Policy Research; Monmouth University Polling Institute; NORC at the University of Chicago; PolicyInteractive; RTI International; University of Illinois at Springfield Survey Research Office; University of Kentucky Survey Research Center; The University of Michigan Survey Research Operations, Survey Research Center; University of Nebraska Bureau of Sociological Research; University of Northern Iowa Center for Social and Behavioral Research; University of Wisconsin Survey Center. In addition, the following organizations provided supporting materials: D3-Designs, Data Decisions; University of North Florida Public Opinion Research Laboratory; University of Wisconsin Survey Center; LAPOP, Vanderbilt University. The task force gratefully acknowledges the information provided by these organizations.

Table 6-1. **Characteristics of SROs responding to task force request for information on policies, guidelines, and procedures**

Category	Number of responding SROs	% of responding SROs
Size of SRO		
≤ 10 permanent employees	6	42.9
11-24 permanent employees	2	14.3
25-99 permanent employees	1	7.1
100-499 permanent employees	1	7.1
500-999 permanent employees	1	7.1
1,000 or more permanent employees	3	21.4
Type of SRO		
Government agency	0	0.0
SRO is part of a college or university	9	64.3
For profit company	2	14.3
Nonprofit company	3	21.4
Have government contracts?		
Yes	8	57.1
No	6	42.9

6.4 Preventing Falsification

Sections of the survey were devoted to items asking about pre-data collection policies and practices, and other organization-level policies and practices not related to the data collection cycle. These items were selected from practices discussed in the literature (see Chapter 3) and largely focused on workplace culture and training.

Workplace Culture and Practices

Several items were on the survey to assess what SROs are doing to promote a positive workplace environment, create a culture that discourages falsification, and have hiring practices that support limiting data falsification and fabrication risks.

The survey results are suggestive of most common and less common approaches:

- Most Common Hiring Practices (≥66.7% of respondents):
  - Use background checks of research staff.
  - Use background checks of field or interviewing staff.
  - Use other screening tools for field or other interviewing staff.
- Less Common Hiring Practices:
  - Use other screening tools for research staff.
  - Hire full-time interviewers.
- Most Common Pay and Benefits Practices (≥66.7% of respondents):
  - Provide time off for family care for research staff.
  - Provide paid sick time off for research staff.
  - Pay interviewers per hour.
  - Provide time off for family care for interviewers and supervisors.
- Less Common Pay and Benefits Practices:
  - Provide paid sick time off for interviewers and supervisors.
  - Pay interviewers bonuses for refusal conversion.
  - Tie research staff promotion opportunities to publication rate.
  - Pay interviewers per complete.
  - Pay interviewers bonuses for completes

In open-ended responses, many organizations also mention requiring all staff, including interviewers, to sign a Pledge of Confidentiality, as well as emphasizing ethical conduct of research repeatedly in interviewer training. Some organizations also pay bonuses or provide sick time off depending on client and survey mode.

Institutional-Level Training

We also asked about training required of interviewers, supervisors, and research staff that would, among other things, discourage falsification. While some SROs noted that there was project-specific variation in human subjects training requirements, most ($\geq$ 66.7%) said they:

- Require human subjects training for research staff, field supervisors, and interviewers; and
- Require other research ethics training for research staff, field supervisors, and interviewers.

In open-ended responses, some SROs indicated that they did not provide interviewers with human subjects training, or only did so if required by contract on specific projects.

6.5 Detecting Falsification

One section of the survey was devoted to items asking about policies and practices during data collection. Another section asked about post-data collection policies and practices. Some of these can be used either during or after data collection. These items were selected from practices discussed in the literature (see Chapter 4) and largely focused on monitoring, verification, and the use of paradata.

During Data Collection

The questionnaire also contained questions on practices regarding monitoring, recording interviews and the use of paradata while the data collection is in progress. The results (see Appendix C) are suggestive of the most common and less common approaches.

- Most Common Monitoring Practices (≥66.7% of respondents):
  - Use supervisors to monitor interviewers in real-time.
  - Monitor interviewer-level unit response rates.
  - Monitor interviewer-level patterns of responses for questions involved in skip patterns.
  - Monitor interviewer-level patterns of responses for questions not involved in skip patterns.
  - Use supervisors to monitor field interviewers during data collection.
  - Less Common Monitoring Practices:
  - Monitor interviewer-level item response rates.
  - Monitor interviewer location at the time of interview using GPS.
  - Use different interviewer monitoring tools for experienced and inexperienced interviewers, vulnerable populations, short field periods, long field periods, and long questionnaires.
- Most Common Uses of Paradata (≥66.7% of respondents):
  - Examine timing paradata related to the length of the questionnaire overall at the interviewer level with the goal of identifying unusual patterns.
  - Examine paradata related to the length of the questionnaire overall for individual respondents with the goal of identifying unusual patterns.
  - Examine interviewer-level patterns of consent for pieces of a study in addition to the questionnaire (e.g., biomarkers, linkage to administrative records).
  - Examine other types of paradata (e.g., backups, answer changes) at the interviewer level with the goal of identifying unusual patterns.

- Less Common Uses of Paradata:
  - Examine other types of paradata (e.g., backups, answer changes) for individual respondents with the goal of identifying unusual patterns.
  - Examine timing paradata related to the length of individual sections in the questionnaire at the interviewer level with the goal of identifying unusual patterns.
  - Examine paradata related to the length of individual sections in the questionnaire for individual respondents with the goal of identifying unusual patterns.

With respect to recording interviews, slightly more than half report recording each interview in total, while fewer use CARI tools to record portions of interviews. Less than half report evaluating signatures on respondent incentive forms. Some organizations report variation across projects in whether interviews are recorded at all and the rate at which interviews are recorded. Some report recording all qualitative interviews while only recording a percentage of quantitative interviews.

Post-Data Collection

Post-data collection items focused on practices involving examining interviewer-level patterns of results and verifying completed interviews.

- Most Common Post-Data Collection Practices (≥66.7% of respondents):
  - Verify interview completion with respondents using telephone calls.
  - Examine interviewer-level patterns of results for biomarker data.
- Less Common Post-Data Collection Practices:
  - Verify interview completion with respondents using in person visits.
  - Verify interview completion with respondents using mailed postcards.

6.6 Evidence of Success

A few organizations note that falsified interviews have been detected through the procedures and algorithms in place, and that interviewer behavior has been monitored and corrected as needed. Others report internal surveys in which employees report integrity as being a key part of the organization. Yet it appears that few formal evaluations, and no experimental evaluations, have been conducted among our SRO respondents to identify the most important or efficacious methods for preventing or detecting falsification attempts.

7. Summary and Recommendations

With heightened public scrutiny of published statistics, data integrity is paramount to the continued viability of survey research. In addition to the public-perception ramifications, data falsification may result in biased estimates, affect the precision of estimates, and affect multivariate relationships reflected in the data. Steps aimed at preventing and detecting falsification expend resources that are often scarce in today's environment. However, failure to establish procedures and policies to prevent and detect falsification could result in much more substantial impacts on the survey, the survey research organization, and the field of survey research as a whole.

Falsification is not a new problem. In early literature on this topic, Crespi (1945) and Bennett (1948) described the "cheater problem" in interviewing. However, technological advances in recent years have resulted in the development of new methods of detection; additionally, increased access to the survey data by a variety of players within survey research organizations has resulted in the need for additional safeguarding of the data against falsification.

Until recently, much of the focus on falsification in the literature has been on "curbstoning," or fabrication of entire interviews by interviewers. However, as noted by Koczela, Furlong, McCarthy, and Mushtaq (2015), falsification goes well beyond interviewer curbstoning. In addition to fabrication of entire interviews or parts of interviews by interviewers, falsification may involve altering of disposition codes, duplicating entire survey data records, or duplicating records with changes to a few answers. A key aspect of falsification is that it is an intentional manipulation of data for the purpose of short-cutting the data collection process or altering the survey findings. Other data-altering activities that are performed for statistically valid purposes using sound and transparent methods, such as imputation or methods used to limit disclosure risk, do not fall under the umbrella of falsification.

Various methods have been developed for preventing and detecting falsification. These include supervisor and interviewer training on the importance of integrity in the data collection effort, reinterview and live monitoring methods, as well as the use of technologies, such as computer-assisted-recorded interviewing (CARI) and global positioning system (GPS) tracking of interviewers to algorithms that analyze patterns in the data to detect duplicates, near-duplicates, and other evidence of fabrication.

Many survey research organizations have developed their own protocols and procedures for preventing, detecting, investigating, and reporting on falsification; however, only recently have best practices for the field been published (American Association for Public Opinion Research (AAPOR 2003; AAPOR 2005). One aim of AAPOR's Transparency Initiative (AAPOR 2017) is to counter falsification.

It is incumbent on survey research organizations to establish policies and practices aimed at preventing falsification and at detecting and removing falsification should it occur. To that end, the AAPOR/ASA Data Falsification Task Force offers the following recommendations:

- Survey research organizations should create an environment where falsification is discouraged and pledge to report it when it is discovered. To that end, survey organizations should work to promote workplace cultures that limit falsification, implement study design features that help minimize the risk of falsification, and ensure that best practices are being followed for training and monitoring of interviewers and research staff before, during, and after data collection.

- Survey research organizations' leadership must clearly articulate to staff the values, goals, and rules of the organization. This applies to all levels of the organization but may be especially important for employees who receive little day-to-day supervision or interaction.

- Survey organizations should regularly reaffirm commitment to the norms and expectations of researchers, interviewers, and field supervisors to 1) collect accurate and reliable data from the selected households and respondents; 2) ask each question in a survey instrument; and 3) ensure that the interviewers and field supervisors are personally responsible for maintaining confidentiality, for visiting the selected sample units, and for reporting any problems that they encounter immediately to the supervisory and managerial staff. Specific training about falsification should be a best practice. In addition to this training, the organization might consider requiring staff to sign a pledge of ethical behavior or a "data integrity" agreement. Survey organizations should also consider conducting internal surveys on the climate of research integrity at all levels.

- To prevent falsification, study-level factors should be taken into account as part of the study design process. Consideration should be given to the risk of falsification associated with the use of long and burdensome questions and questionnaire. Study protocols should be kept as simple as possible. When possible, procedures should be put in place to monitor data collection. Studies should consider the risk of falsification associated with conducting interviews in dangerous or otherwise difficult areas and provide support or alternatives to the data collectors in such situations. Studies should also acknowledge and consider the risk of falsification associated with large cash incentives.

- Survey organizations should use methods aimed at preventing falsification during data collection by reducing the perceived opportunity to falsify data without being caught. These methods include monitoring, use of CARI, analysis of paradata, use of verification interviews (reinterviews), and real-time use of GPS devices (e.g., geofencing).

- To the extent possible, methods for detection of falsification should be used in combination to achieve an optimal balance of accurate detection and cost. Not every method or technology is available or applicable to every situation. Some have significant costs but can yield a high level of confidence in the results. Others may be cheaper or quicker to implement but may miss important potential sources of falsification. When considering and applying output monitoring methods of detection, organizations should bear in mind that these methods cannot lead to a "confirmation" of falsification. If a data point or pattern does not fit the researcher's expectation based on knowledge of the survey, respondents, or the subject matter, falsification may be one of several possible explanations. Other explanations could include measurement error due to reasons other than falsification, including unintentional interviewer errors, respondent confusion, satisficing behavior, etc. If it is determined that an individual has falsified data, personnel actions, from reprimands to dismissal, should be initiated.

- Survey research organizations should be encouraged to provide proper documentation of the falsification detection process and its impact on data quality.

- Survey research organizations should be encouraged to publish their successes and failures with detecting falsification, even if this requires anonymization of the specific survey and indexing actual data. This will help other organizations learn from these experiences.

References

American Association for Public Opinion Research. (2003). Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf on March 17, 2018.

American Association for Public Opinion Research. (2005). Policy Regarding Certification of Survey Research Organizations for Interviewer Falsification Protocols. Retrieved from https://www.aapor.org/Standards-Ethics/Resources/Interviewer-Falsification-Practices-and-Policies/Falsification-Certification.aspx on May 3, 2019.

American Association for Public Opinion Research. (2009). Report to the Standards Committee on the status of Human Subjects Protection Training Requirements. Retrieved from www.aapor.org/Education-Resources/Reports/Status-of-Human-Subjects-Protection-Training-Requi.aspx

American Association for Public Opinion Research. (2017). AAPOR Transparency Initiative: AAPOR Policies and Procedures for Transparency Certification. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/TI-operational-procedures-10-4-17.pdf.

Antes, A. L., Mart, A., & DuBois, J. M. (2016). Are leadership and management essential for good research? An interview study of genetic researchers. *Journal of Empirical Research on Human Research Ethics, 11*, 408-423.

Babbage, C. (1830). Reflections on the decline of science in England and on some of its causes. In M. Campbell-Kelly (Ed.),*The works of Charles Babbage*. London, England: Andesite Press.

Barney, J. B. (1986). Organizational culture: Can it be a source of sustained competitive advantage? *The Academy of Management Review, 11*(3), 656-665. Retrieved from http://www.jstor.org/stable/258317.

Bates, N., Dahlhamer, J., Phipps, P., Safir, A., & Tan, L. (2010). Assessing contact history paradata quality across several federal surveys. *Proceedings of the Joint Statistical Meetings.*

Bedeian, A. G., Taylor, S. G., Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning and Education*, *9*(4), 715-725.

Bennett, A. S. (1948). Toward a solution of the "cheater problem" among part-time research investigators. *Journal of Marketing*, *12*(4), 470-474. doi: 10.2307/1246628

Biemer, P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 147-168.

Biemer, P., Herget, D., Morton, J., & Willis, G. (2000). The feasibility of monitoring field interview performance using computer audio-recorded interviewing (CARI). *Proceedings of the Joint Statistical Meetings*, 1068-1073.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: John Wiley & Sons.

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.

Biemer, P., & Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. In L. E. Lyberg, P. P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality*, 601-632. Hoboken, NJ: John Wiley & Sons.

Birnbaum, B., DeRenzi, B., Flaxman, A. D., & Lesh, N. (2012, March 11-12). *Automated quality control for mobile data collection.* Presented at the Second ACM Symposium on Computing for Development, Atlanta, GA.

Blasius, J., & Thiessen, V. (2012). Institutional quality control practices. In J. Blasius & V. Thiessen (Eds.), *Assessing the Quality of Survey Data* 57-80. London: Sage.

Blasius, J., & Thiessen, V. (2013). Detecting poorly conducted interviews. In P. Winker, N. Menold, & R. Porst (Eds), *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention* (pp. 67-88). Frankfurt: Peter Lang.

Blasius, J., & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research*, *52*, 479-493.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. Heckman & E. Learners (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3705-3843).

Boyd, H. W., & Westfall, R. (1955). Interviewers as a source of error in surveys. *Journal of Marketing, 19*(4), 311-324. doi: 10.2307/1247046.

Bredl, S., Winker, P., & Kötschau, K. (2008). *A statistical approach to detect cheating interviewers.* Discussion Paper, No. 39, Justus-Liebig-Universität Gießen, Zentrum für Internationale Entwicklungs und Umweltforschung (ZEU), Giessen.

Bredl, S., Storfinger, N., & Menold, N. (2011). *A literature review of methods to detect fabricated survey data.* Discussion Paper No. 56, Center for International Development and Environmental Research, Justus Liebig University Giessen.

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

Breiman, L. (1994). The 1991 census adjustment: Undercount or bad data? *Statistical Science, 9*(4), 458-475. doi: 10.1214/ss/1177010259.

Broockman, D., Kalla, J., & Aronow, P. (2015). Irregularities in LaCour (2014). Retrieved from http://stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf.

Bruderl, J., Huyer-May, B., & Schmiedeberg, C. (2013). Interviewer behavior and the quality of social network data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewer deviations in surveys* (pp. 157-160).

Burke, B. (2015). *No address, no problem: The innovative use of technology to accurately complete a building survey in the Emirate of Abu Dhabi*. Presented at the World Association for Public Opinion Research 2015 Regional Conference. Qatar University, Doha.

Burnham, G., Lafta, R., Doocy, S., & Roberts, L. (2006). Mortality after the 2003 invasion of Iraq: A cross-sectional cluster sample survey. *The Lancet*, *368*(9545), 1421-1428.

Bushery, J. M., Reichert, J. W., Albright, K .A., & Rossiter, J. C. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the Joint Statistical Meetings,* 316-320.

Buzas, J. S., Stefanski, L. A., & Tosteson, T. D. (2005). Measurement error. In W. Ahrens & I. Pigeot (Eds.), *Handbook of epidemiology* (pp. 1241-1282). Springer: Berlin.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*. Boston, MA: Springer.

Chai, J. J. (1971). Correlated measurement errors and the least squares estimator of the regression coefficient. *Journal of the American Statistical Association, 66*(335), 478-483.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*(4), 637.

Converse, P. E. (2006). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent*. New York: The Free Press of Glencoe.

Crespi, L. P. (1945). The cheater problem in polling. *The Public Opinion Quarterly, 9*(4), 431-445. doi: 10.2307/2745558

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

Dahlhamer, J. M., Taylor, B., Simile, C. M., & Stussman, B. J. (2009). *Using paradata to assess and monitor data quality in the National Health Interview Survey (NHIS).* Presentation given at the U.S. Census Bureau Sponsors Meeting, Suitland, MD, January 13.

Daikeler, J., Silber, H., Bosnjak, M., Zabal, A., & Martin, S. (2017). *A general interviewer training curriculum for computer-assisted personal interviews* (GIT-CAPI; Version 1, 2017). GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz-Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_022

Daily Research News Online. (2016). JIR group wins respondent data falsification case. Retrieved from http://www.mrweb.com/drno/news22890.htm

Dajani, A. N., & Marquette, R. J. (2015). *Reinterview detection and prevention at Census: New initiatives*. Paper presented at the Curbstoning Seminar Part III, Washington Statistical Society, Washington, DC.

Davis, M. S., Riske-Morris, M., & Diaz, S. R. (2007). Causal factors implicated in research misconduct: Evidence from ORI case files. *Science and Engineering Ethics, 13*(4), 395-414. doi: 10.1007/s11948-007-9045-2

DeGeest, D. S., Follmer, E. H., Walter, S. L., & O'Boyle, E. H. (2015). The benefits of benefits: A dynamic approach to motivation-enhancing human resource practices and entrepreneurial survival. *Journal of Management*, *43*(7): 2303-2332. RETRACTED.

DeGeest, D. S., Follmer, E. H., & Lanivich, S. E. (2016). Timing matters: When high-performance work practices enable new venture growth and productivity. *Journal of Management*, *44*(4), NP6-NP33. RETRACTED.

Dellaportas, S. (2013). Conversations with inmate accountants: Motivation, opportunity and the fraud triangle. *Accounting Forum, 37*(1), 29-39.

Dorminey, J., Fleming, A. S., Kranacher, M.-J., & Riley, R. A., Jr. (2012). The evolution of fraud theory. *Issues in Accounting Education, 27*(2), 555-579.

DuBois, J. M., & Antes, A. L. (2018). Five dimensions of research ethics: A stakeholder framework for creating a climate of research integrity. *Academic Medicine 93*(4), 550-555.. doi:10.1097/acm.0000000000001966

DuBois, J. M., Chibnall, J. T., Tait, R. C., & Vander Wal, J. S. (2016). Lessons from researcher rehab. *Nature, 534,* 173-175.

Eckman, S., & Kreuter, F. (2013). Undercoverage rates and undercoverage bias in traditional housing unit listing. *Sociological Methods & Research, 42*(3), 264-293. doi:10.1177/0049124113500477

Edwards, B., & Maitland, A. (2015). *Quantifying measurement error.* Paper presented at the 6th conference of the European Survey Research Association, Reykjavik, Iceland, July 14-17, 2015.

Edwards, B., Maitland, A., & Connor, S. (2017). Measurement error in survey operations management: detection, quantification, visualization, and reduction. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 255-277). New York: John Wiley & Sons.

Ellis, C., Sikes, N., Sage, A., Eyerman, J., & Burke, B. (2011). *Technological advances to reduce survey error.* Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Eng, J. L. V., Wolkon, A., Frolov, A. S., Terlouw, D. J., Eliades, M. J., Morgah, K., … Hightower, A .W. (2007). Use of hand-held computers with global positioning systems for probability sampling and data entry in household surveys. *The American Journal of Tropical Medicine and Hygiene*, *77*(2), 393-399.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738, 1-11.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS ONE, 5*(4), e10271. doi: 10.1371/journal.pone.0010271

Faranda, R. (2015). *The cheater problem revisited: Lessons from six decades of State Department polling.* Paper presented at New Frontiers in Preventing, Detecting, and Remediating Fabrication in Survey Research conference, NEAAPOR, Cambridge, MA.

Finn, A., & Ranchhod, V. (2017). Genuine fakes: The prevalence and implications of data fabrication in a large South African survey. *The World Bank Economic Review*, *31*(1), 129-157. doi:10.1093/wber/lhv054

Fischer, M., West, B. T., Elliott, M. R., & Kreuter, F. (2018). The impact of interviewer effects on regression coefficients. *Journal of Survey Statistics and Methodology.* Retrieved from https://doi.org/10.1093/jssam/smy007

Forsman, G., & Schreiner, I. (1991). The design and analysis of reinterview: An overview. In P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*, Chapter 15. New York: John Wiley & Sons.

Fuller, W. (1987). *Measurement error models.* Wayne A. Fuller (Ed.). Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.

George, S. L., & Buyse, M. (2015). Data fraud in clinical trials. *Clinical investigation, 5*(2), 161-173. doi:10.4155/cli.14.116

Giles, J. (2007). Death toll in Iraq: Survey team takes on its critics. *Nature*, *446*(7131), 6-7.

Goldenring, J. R. (2010). Innocence and due diligence: Managing unfounded allegations of scientific misconduct. *Academic Medicine, 85*(3), 527-530.

Groves, R. M. (1989). *Survey errors and survey costs.* Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.

Gwartney, P. A. (2013). Mischief versus mistakes: Motivating interviewers to not deviate. In P. Winkler, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys: Impact, reasons, detection, and prevention* (pp. 195-216). Frankfurt am Main, Germany: PL Academic Research.

Hansen, M. H., Hurwitz, W. N., & Bershad, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute, 38*(2), 359-374.

Harding, D., & Jackson, P. (2012). *Quality in market research: From theory to practice*. London: BSI Standards Limited.

Harrison, D. E., & Krauss, S. I. (2002). Interviewer cheating: Implications for research on entrepreneurship in Africa. *Journal of Developmental Entrepreneurship*, *7*(3), 319-330.

Hauck, M. (1969). Is survey postcard verification effective? *Public Opinion Quarterly*, *33*(1), 117-120.

Hicks, W. D., Edwards, B., Tourangeau, R., McBride, B., Harris-Kojetin, L. D., & Moss, A. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, *74,* 985-1003.

Hill, T.P. (1995). A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, *10*(4), 354–363.

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

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons.

Inciardi, J. A. (1981). Ficticious data in drug abuse research. *International Journal of the Addictions, 16*(2), 377-380.

Institute of Medicine and National Research Council. (2002). *Integrity in scientific research: Creating an environment that promotes responsible conduct*. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17225/10430.

Jans, M., Sirkis, R., Schultheis, C., Gindi, R., & Dahlhamer, J. (2011). Comparing CAPI trace file data and quality control reinterview data as methods of maintaining data quality. *Proceedings of the Joint Statistical Meetings*, 477-489.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-32.

Johnson, N. F., Spagat, M., Gourley, S., Onnela, J.-P., & Reinert, G. (2008). Bias in epidemiological studies of conflict mortality. *Journal of Peace Research, 45*(5), 653-663. doi:10.1177/0022343308094325

Johnson, T. P., Parker, V., & Clements, C. (2001). Detection and prevention of data falsification in survey research. *Survey Research: Newsletter from the Survey Research Laboratory*, *32*(3), 1-2.

Jones, L. D. (Ed.). (1986). Philosophy and principles of data analysis 1965-1986. In *The Collected Works of John W. Tukey, Volume IV*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Josten, M., & Trappmann, M. (2016). Interviewer effects on a network-size filter question. *Journal of Official Statistics, 32*(2), 349-373.

Keating, M., Loftis, C., McMichael, J., & Ridenhour, J. (2014). *New dimensions of mobile data quality.* Paper presented at the Federal CASIC Workshops, Washington, DC.

Kemper, C .J., & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10*(3), 92–99. doi: 10.1027/1614-2241/a000078

Kennickell, A. (2015). Curbstoning and culture. *Statistical Journal of the IAOS, 31,* 237-240.

Kindred, G. M., & Scott, J. B. (1993). Fabrication during the 1990 Census nonresponse followup. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 335-340.

Kingori, P., & Gerrets, R. (2016). Morals, morale and motivations in data fabrication: Medical research fieldworkers views and practices in two sub-Saharan African contexts. *Social Science & Medicine, 166,* 150-159. doi: https://doi.org/10.1016/j.socscimed.2016.08.019

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association, 57*(297), 92-115.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

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

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). Curbstoning and beyond: Confronting data fabrication in survey research. *Statistical Journal of the IAOS*, *31*(3), 413-422.

Kosyakova, Y., Skopek, J., & 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. doi:10.1093/ijpor/edu027

Krejsa, E., Davis, M., & Hill, J. (1999). Evaluation of the quality assurance falsification interview used in the Census 2000 dress rehearsal. *Proceedings from Section on Survey Research Methods*, 635-640.

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research, 40*(1), 88-104. doi: 10.1177/0049124110392342

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS, 32,* 283-291.

LaCour, M. J., & Green, D. P. (2014). When contact changes minds: An experiment on transmission of support for gay equality. *Science*, *346*(6215), 1366-1369. RETRACTED.

Landrock, U. (2017). Investigating interviewer falsifications – A quasi-experimental design. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 136*(1), 5-20.

Li, J., Brick, J. M., Tran, B., & Singer, P. (2011). Using statistical models for sample design of a re-interview program. *Journal of Official Statistics, 27*(3), 433-450.

Mahalanobis, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society, 109,* 325-370.

McCook, A. (2018). A cancer researcher said she collected blood from 98 people. It was all her own. *Retraction Watch*, May 14.

McGonagle, K. A., Brown, C., & Schoeni, R. F. (2015). The effects of respondents' consent to be recorded on interview length and data quality in a national panel study. *Field Methods*, *27*(4), 373-390. https://doi.org/10.1177/1525822X15569017

Marker, D. A. (2008). Review: Methodological review of "Mortality after the 2003 invasion of Iraq: A cross-sectional cluster sample survey." *Public Opinion Quarterly, 72*(2), 345-363. doi:10.1093/poq/nfn009

Menold, N., & Kemper, C.J. (2014). How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys. *International Journal of Public Opinion Research*, *26*(1), 41-65.

Mochty, L. (2002). Die Aufdeckung von Manipulationen im Rechnungswesen: Was leistet das Benford's Law. *Die Wirtschaftsprufung, 55*(14), 725-736.

Mulry, M. H., & Spencer, B. D. (1991). Total error in PES estimates of population. *Journal of the American Statistical Association, 86*(416), 839-855. doi:10.2307/2290495

Murphy, J. J., Baxter, R. K., Eyerman, J. D., Cunningham, D., & Kennet, J. (2004, May 13-16). *A system for detecting interviewer falsification.* Presented at the American Association for Public Opinion Research 59th Annual Conference, Phoenix, AZ.

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

Murphy, J. J., Eyerman, J. D., McCue, C., Hottinger, C., & Kennet, J. (2005, October 25-28*).* Interviewer falsification detection using data mining. *Proceedings of Statistics Canada Symposium 2005 Methodological Challenges for Future Information Needs.* Ottawa, ON.

National Academy of Sciences. (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age.* Washington, DC: National Academies Press. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK215265/

National Academies of Sciences, Engineering, and Medicine (2009). *Ensuring the integrity, accessibility, and stewardship of research data in the digital age.* Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2017). *Fostering integrity in research.* Washington, DC: National Academies Press.

National Research Council & Institute of Medicine. (2002). *Integrity in scientific research: Creating an environment that promotes responsible conduct*. Committee on Assessing Integrity in Research Environments. Washington, DC: National Academies Press.

Neuhaus, J. M., & Segal, M. R. (1993). Design effects for binary regression models fitted to dependent data. *Statistics in Medicine, 12*(13), 1259-1268.

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

Office of Research Integrity. (2002). Can survey research staff commit scientific misconduct? *ORI Newsletter*, *10*(3). Retrieved from https://ori.hhs.gov/images/ddblock/vol10_no3.pdf

Office of Science and Technology Policy. (2000, December 6). Federal Research Misconduct Policy. *Federal Register, 65*(235), 76260-76264.

O'Muircheartaigh, C. (1977). Response errors. In C. O'Muircheartaigh, & C. D. Payne (Eds.), *The analysis of survey data*. London; New York: Wiley.

Parsons, J. (2016). *Preventing and detecting interviewer falsification*. Presentation given at a University of Illinois-Chicago survey research laboratory seminar.

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

Rassler, S., Rubin, D. B., and Zell, E. R. (2011). Imputation. In P. J. Lavrakas (Ed.), *Encyclopedia of Research Methods*. Thousand Oaks, CA: Sage Publications.

Robbins, M. (2015, February). Preventing data falsification in survey research: Lessons from the Arab barometer. In *New Frontiers in Preventing, Detecting, and Remediating Fabrication in Survey Research. Conferences hosted by NEAAPOR and Harvard Program on Survey Research, Cambridge, MA* (Vol. 13).

Robbins, M. (forthcoming, 2019). New frontiers in detecting data fabrication. In T. P. Johnson, B. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology* (1st Ed.). John Wiley & Sons.

Robbins, M., & Noble, K. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*(3), 283-291.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., & Wagner, G. G. (2005). Automatic identification of faked and fraudulent interviews in surveys by two different methods. *Proceedings of the Joint Statistical Meetings, American Statistical Association, Survey Research Method Section.* Toronto, ON.

Schnell, R. (1991). Der Einfluß gef¨alschter interviews auf survey-ergebnisse. *Zeitschrift f¨ur Soziologie*, *20*(1), 25-35.

Schräpler, J.-P. (2010). Benford's Law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP). *SOEP papers on Multidisciplinary Panel Data Research*, No. 273, Deutsches Institut für Wirtschaftsforschung-(DIW), Berlin

Schraepler, J.-P., & Wagner, G. (2005). Identification, 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.

Schreiner, I., Pennie, K., & Newbrough, J. (1988). Interviewer falsification in Census Bureau surveys. *Proceedings from Section on Survey Research Methods*, 491-496.

Scott, P. and Fasli, M. (2001). Benford's Law: An empirical investigation and a novel explanation. *CSM Technical Report 349*. Department of Computer Science, University of Essex.

Sharma, S. (2016). *Detecting falsification in an audience measurement survey.* NORC conference on New Approaches to Dealing with Survey Data Fabrication, Bethesda, MD..

Sharma, S., & Elliott, M. R (2019, September 19). Detecting falsification in a television audience measurement panel survey. *International Journal of Market Research*. https://doi.org/10.1177/1470785319874688

Silver, N. (2009, September 26). Comparison study: Unusual patterns in strategic vision polling data remain unexplained. Retrieved from https://fivethirtyeight.com/features/comparison-study-unusual-patterns-in/ Accessed August 26, 2017.

Simmons, K., Mercer, A., Schwarzer, S., & Kennedy, C. (2016). Evaluating a new proposal for detecting data falsification in surveys. *Statistical Journal of the IAOS, 32,* 327-338.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875-1888.

Smith, P. B., MacQuarrie, C. R., Herbert, R. J., Cairns, D. L., & Begley, L. H. (2004). Preventing data fabrication in telephone survey research. *Journal of Research Administration, 35*(2), 13-20.

Smith, T. W., & Sokolowski, J. (2008). *Using audio-visuals in surveys.* NORC/University of Chicago: GSS Methodological Report No. 112.

Spagat, M. (2010). Ethical and data-integrity problems in the second *Lancet* survey of mortality in Iraq. *Defence and Peace Economics, 21*(1), 1-41.

Spagat, M. (n.d.). *Suspicious supervisors and suspect surveys.* Stats.org. Retrieved from https://senseaboutscienceusa.org/suspicious-supervisors-suspect-surveys/

Spagat, M., & Dougherty, J. (2010, May). Conflict deaths in Iraq: A methodological critique of the ORB survey estimate. In *Survey Research Methods 4*(1), 3-15.Stern, V. (2018). Management researcher admits to falsification, resigns. *Retraction Watch*, March 21.

Stringer, C., & Dahlhamer, J. (2010). *PANDA: Using paradata for quality and performance monitoring.* Presented at the U.S. Census Bureau Workshop: Advances in Responsive and Adaptive Survey Designs, Suitland, MD, October 14-15.

Sukhatme, P. V. & Seth, G. R. (1952). Nonsampling errors in surveys. *Journal of Indian Society of Agricultural Statistics*, 5, 5-41

Swanson, D., Cho, M. J., & Eltinge, J. (2003), Detecting possibly fraudulent or error-prone survey data using Benford's Law. *Proceedings of the Section on Survey Research Methods,, American Statistical Association, Survey Research Method Section,* 4172-4177.

Thissen M. R., & Myers, S. K. (2016). Systems and processes for detecting interviewer falsification and assuring data collection quality. *Statistical Journal of the IAOS*, *32,* 339-347.

Thissen, M. R., & Rodriguez, G. (2004). Recording interview sound bites through Blaise instruments. *Proceedings of the International Blaise Users' Conference,* 411-423.

Titus, S. L., Wells, J. A., & Rhoades, L. J. (2008). Repairing research integrity. *Nature*, *453*(7198), 980.

Tukey, J. W. (1977). *Exploratory data analysis*. New York: Addison-Wesley.

Turner, C. F., Gribble, J. N., Al-Tayyib, A. A., & Chromy, J. R. (2002). Falsification in epidemiological surveys: Detection and remediation. *Technical Papers on Health and Behavior Measurement, No. 53.* Washington, DC: Research Triangle Institute.

Tremblay, A. (1991). *Evaluation Project P5: Analysis of PES P-sample fabrications from PES quality control data.* 1990 Coverage Studies and Evaluation Memorandum Series #E-4. Washington, DC: U.S. Census Bureau.

U.S. Bureau of the Census (1985). *Evaluation of censuses of population and housing.* Washington, DC: Author.

U.S. Department of Health and Human Services (2005). *Public Health Service Policies on Research Misconduct,* 42 CFR Parts 50 and 93, Section 93.103 (Available at https://ori.hhs.gov/sites/default/files/42_cfr_parts_50_and_93_2005.pdf, page 28386; accessed January 5, 2021).

Valente, T. W., Dougherty, L., & Stammer, E. (2017). Response bias over time: Interviewer learning and missing data in egocentric network surveys. *Field Methods, 29*(4), 303-316. doi:10.1177/1525822x17703718

Wagner, J., Olson, K., & Edgar, M. (2017, October). The utility of GPS data in assessing interviewer travel behavior and errors in level-of-effort paradata. *Survey Research Methods*, *11*(3), 218-233.

Watts, L. L., Medeiros, K. E., Mulhearn, T. J., Steele, L. M., Connelly, S., & Mumford, M.D. (2017). Are ethics training programs improving? A meta-analytic review of past and present ethics instruction in the sciences. *Ethics & Behavior, 27*(5), 351-384. doi: 10.1080/10508422.2016.1182025

Wells, J. T. (2017). *Corporate fraud handbook: Prevention and detection* (5th Ed.). Hoboken, NJ: John Wiley & Sons.

West, B. T., & Blom, A. G. (2016). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology, 5*(2), 175-211.

Wetzel, A. (2003). Assessing the effect of different instrument modes on reinterview results from the Consumer Expenditure Quarterly Interview Survey. *Proceedings of the American Statistical Association (Survey Research Methods Section)* (pp. 4508-4513).

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.

Yi, G. Y. (2017). *Statistical analysis with measurement error or misclassification*. Springer Series in Statistics. New York: Springer New York.

APPENDIX A: Letter Used for SRO Request for Guidelines

Dear Colleague,

We are contacting you on behalf of the joint AAPOR/ASA Data Falsification Task Force (DFTF) with a request for information. Our charge is to develop a white paper on best practices to prevent, detect, and when necessary ameliorate the effects of data falsification and fabrication by survey and public opinion organization personnel, including but not limited to interviewers, supervisors, and researchers.

In particular, we would like to include in our report a review of the current landscape. The DFTF is hoping to collect information on existing organization guidelines and policies in our profession regarding data falsification and fabrication. We are asking that you share any written guidelines or technical documents your organization may have developed in this regard. We are interested in policies and practices with respect to: pre-data collection such as new interviewer and human subjects protection training, signed data integrity affidavits, interviewer and others' awareness of monitoring and verification practices; real time data collection monitoring, dashboards, paradata review; and post-data collection such as statistical modeling tools. To assist you we have developed a brief survey with a checklist of things organizations may do to discourage and detect data falsification and fabrication.

We are also interested in any evidence you have – documented or anecdotal – that these policies or guidelines are effective. This can include any research or case studies you've performed to assess the effectiveness of these approaches. If your organization does not have formal written guidelines or policies regarding data falsification, that information is valuable to us as well.

Any information you share with us will be anonymized in our report. Our goal is to provide a report that will assist survey and public opinion research organizations in strengthening their capacity for falsification prevention and detection by integrating best practices and insights gained from the experiences of researchers like yourself.

If you are willing to share any additional documentation or guidelines, please send them to Ron Langley at langley@uky.edu by **May 11**. Alternatively you can send the materials to the DFTF co-chairs below.

Thank you for considering this request! We will be happy to acknowledge organizations that provide us information in the final report. If you have any questions, feel free to contact us. If you are not the best person in your organization to respond to this request, please forward it to the appropriate person.

Best Regards,
Jill DeMatteis, JillDematteis@westat.com
Linda Young, Linda.Young@nass.usda.gov
Task Force Co-chairs

APPENDIX B: Questionnaire Used for SRO Request for Guidelines

AAPOR/ASA Data Falsification Task Force Survey of Survey Research Organizations

The DFTF has put together the following checklist of things organizations may do to prevent and detect data falsification and fabrication. The survey is loosely organized by phases of data collection – pre, during, and post. At the end of each section there is a text box if you wish to elaborate upon any of your answers, or tell us of things your organization does that we neglected to ask about. It is possible that your organization does some of these things only for certain projects and not others. If so, please answer yes to those items. You can use the text box to elaborate if you wish.

Please note that you can stop at any time and return and pick up where you left off by scrolling to the end of the survey and clicking the 'save and return later' button.

THANK YOU for agreeing to help us out!

Organizations do a variety of things to train staff and discourage or identify data fabrication and falsification. Please indicate whether you use each of the following at your organization. Space is provided after each section you if you wish to elaborate.

Pre-Data Collection Policies and Practices

Does your organization conduct background checks before hiring field or interviewing staff?
○ Yes ○ N ○ N/

Does your organization use other screening tools before hiring field or interviewing staff?

○ Yes ○ N ○ N/

Does your organization conduct background checks before hiring other research staff (programmers, coders, investigators, etc.)?
○ Yes ○ N ○ N/

Does your organization use other screening tools before hiring other research staff (programmers, coders, investigators, etc.)?

○ Yes ○ N ○ N/

Does your organization require human subjects training for interviewers?

○ Yes ○ N ○ N/

Does your organization require other research ethics training for interviewers?

○ Yes ○ N ○ N/

Does your organization require human subjects training for field supervisors?

○ Yes ○ N ○ N/

Does your organization require other research ethics training for field supervisors?

○ Yes ○ N ○ N/

Does your organization require human subjects training for other research staff (programmers, coders, investigators, etc.)?

○ Yes ○ N ○ N/

Does your organization require other research ethics training for other research staff (programmers, coders, investigators, etc.)?

○ Yes ○ N ○ N/

Additional comments about pre-data collection policies and practices?

Policies and Practices During Data Collection

Does your organization monitor interviewer-level UNIT response rates?

○ Ye ○ N ○ N

Does your organization monitor interviewer-level ITEM

○ Ye ○ N ○ N

Does your organization monitor interviewer-level patterns of responses for questions involved

○ Ye ○ N ○ N

Does your organization monitor interviewer-level patterns of responses for questions NOT involved in skip patterns?

○ Ye ○ N ○ N/

Does your organization monitor interviewer location at the time of interview using

○ Ye ○ N ○ N/

Does your organization use supervisors to monitor telephone interviewers in real time?

○ Ye ○ N ○ N/

Does your organization use supervisors to monitor field interviewers during data collection?

○ Ye ○ N ○ N/

Does your organization use different interviewer monitoring tools for vulnerable populations?

○ Ye ○ N ○ N/

Does your organization use different interviewer monitoring tools for long questionnaires?

○Y ○N ○N

Does your organization use different interviewer monitoring tools for

○Y ○N ○N

Does your organization use different interviewer monitoring tools for

○Y ○N ○N

Does your organization use different interviewer monitoring tools for experienced and

○Y ○N ○N

Does your organization audio record each

○Y ○N ○N

Does your organization use Computer-Assisted Recorded Interviewing tools to

○Y ○N ○N

Does your organization use in person visits to verify interview

○Y ○N ○N

Does your organization use telephone calls to verify interview

○Y ○N ○N

Does your organization use mailed postcards to verify interview

○Y ○N ○N

Additional comments about policies and practices during data collection?

Post-Data Collection Policies and Practices (may also be using during data collection)

Does your organization examine timing paradata related to the length of the questionnaire overall at the interviewer level with the goal of identifying unusual patterns?

○ Ye ○ N ○ N/

Does your organization examine timing paradata related to the length of individual sections in the questionnaire at the interviewer level with the goal of identifying unusual patterns?

○ Yes ○ No ○ N/A

Does your organization examine other types of paradata (e.g., backups, answer changes) at the interviewer level with the goal of identifying unusual patterns?

○ Y ○ N ○ N/

Does your organization examine paradata related to the length of the questionnaire overall for individual respondents with the goal of identifying unusual patterns?

○ Ye ○ N ○ N/

Does your organization examine paradata related to the length of individual sections in the questionnaire for individual respondents with the goal of identifying unusual patterns?

○ Ye ○ N ○ N/

Does your organization examine other types of paradata (e.g., backups, answer changes) for individual respondents with the goal of identifying unusual patterns?

○ Ye ○ N ○ N/

Does your organization examine interviewer-level patterns of consent for pieces of a study in addition to the questionnaire (e.g., biomarkers, linkage to administrative records)?

○ Ye ○ N ○ N/

Does your organization examine interviewer-level patterns of results for biomarker data?

○ Ye ○ N ○ N/

Does your organization evaluate signatures on respondent incentive forms for authenticity?

○ Ye ○ N ○ N/

Additional comments about post-data collection policies and practices?

Other Organization Level Policies and Practices

Does your organization provide time off for interviewers and supervisors for family care?

○ Ye ○ N ○ N/

Does your organization provide time off for other research staff (programmers, coders, supervisors, monitors, investigators, etc.) for family care?

○ Ye ○ N ○ N/

Does your organization provide paid sick leave time off for interviewers and supervisors?

○ Ye ○ N ○ N/

Does your organization provide paid sick leave time off for other research staff (programmers, coders, supervisors, monitors, investigators, etc.)?

○ Ye ○ N ○ N/

Does your organization tie the promotion opportunities for research staff to their rate of publishing in peer-reviewed journals?

○ Ye ○ N ○ N/

Does your organization hire full time interviewers?

○ Ye ○ N ○ N/

Does your organization pay interviewers per completed interview?

○ Ye ○ N ○ N/

Does your organization pay interviewers per hour?

○ Ye ○ N ○ N/

Does your organization pay interviewers bonuses for completes?

○ Ye ○ N ○ N/

Does your organization pay interviewers bonuses for refusal conversion?

○ Ye ○ N ○ N/

Additional comments on other organizational policies and practices?

Please describe the TYPES OF TRAINING that your organization uses to encourage ethical research practices or discourage data falsification and fabrication.

If you are able to share your training documentation, the committee would appreciate seeing it. Please email materials to langley@uky.edu.

Please describe the TYPES OF PRACTICES that your organization uses to encourage ethical research practices or discourage data falsification and fabrication.

If you are able to share your documentation about any types of practices that you use, the committee would appreciate seeing it. Please email materials to langley@uky.edu.

Please describe what your organization does to CREATE A CULTURE that encourages ethical research practices or discourages data falsification and fabrication.

If you are able to share documentation about company culture practices, the committee would appreciate seeing it. Please email materials to langley@uky.edu.

Please describe the results of any research or case studies your organization has conducted to ASSESS THE EFFECTIVENESS of the training and practices to prevent and detect falsification.

If you are able to share these studies, the committee would appreciate seeing them. Please email materials to langley@uky.edu.

Please tell us which of the following best describes your organization:

○ Academic Institution
○ Government Agency
○ For-Profit Private organization
○ Non-Profit Private Organization
○ Other

Please tell us how many full-time regular employees work for your organization:

○ 1 - 10
○ 11 -24
○ 25 -99
○ 100 - 499
○ 500 - 999
○ 1000+

Does your organization have any federal contracts?

○ Ye
○ N
○ N/

Thank you for your participation in this survey!

It is possible that the Task Force may want to follow up with some respondents about their answers. If you are willing to be contacted, please provide your e-mail address below.

____

APPENDIX C: Survey Methods and Results

METHODS

The DFTF survey was conducted by the University of Kentucky Survey Research Center as a courtesy to AAPOR and ASA and was not funded by any other entity. As discussed in Section 3 the sampling frame was the April 2018 list of AAPOR Transparency Initiative members supplemented by a few other large organizations who were not TI members. The survey was administered in English as a web survey using REDCap. Data were collected from April 23 through June 30, 2018. Requests were sent to 78 SRO's on April 23rd with follow-up requests to non-respondents on May 7th and again on June 8th. Fourteen SRO's completed the interview and 2 indicated they were not eligible because they did not collect data (but contracted it out to others in the sample). This was not a random sample and as such no measures of precision are provided. The results are solely intended to inform what some prominent SRO's are doing to prevent and detect falsification. The request letter is provided in Appendix A and the questionnaire is provided in Appendix B. The results to the survey are presented below with the exception of responses to the open ended questions. Those responses are summarized in Section 3 and are available upon request by contacting Ron Langley at langley@uky.edu.

RESULTS

Pre-Data Collection Policies and Practices

Does your organization:	Yes	No	N/A	Valid %
Conduct background checks before hiring field or interviewing staff	12	1	1	92.3
Use other screening tools before hiring field or interviewing staff	9	4	1	69.2
Conduct background checks before hiring research staff	14	0	0	100.0
Use other screening tools before hiring research staff	8	5	1	61.5
Require human subjects training for interviewers	10	4	0	71.4
Require other research ethics training for interviewers	12	2	0	85.7
Require human subjects training for field supervisors	11	2	1	84.6
Require other research ethics training for field supervisors	12	1	1	92.3
Require human subjects training for research staff	12	2	0	85.7
Require other research ethics training for research staff	12	1	1	92.3

Policies and Practices During Data Collection

Does your organization:	Yes	No	N/A	Valid %
Monitor interviewer-level unit response rates	11	1	2	91.7
Monitor interviewer-level item response rates	6	6	2	50.0
Monitor patterns of responses for questions involved in skip patterns	8	4	2	75.0
Monitor patterns of responses for questions not involved in skip patterns	8	4	2	75.0
Monitor interviewer location at the time of interview using GPS	3	6	4	33.3
Use supervisors to monitor telephone interviewers in real time	12	1	0	92.3
Use supervisors to monitor field interviewers during data collection	6	1	6	85.7
Use different interviewer monitoring tools for vulnerable populations	5	6	2	45.5
Use different interviewer monitoring tools for long questionnaires	4	9	0	30.8
Use different interviewer monitoring tools for long field periods	4	7	2	36.4
Use different interviewer monitoring tools for short field periods	4	7	2	36.4
Use different interviewer monitoring tools for experienced and inexperienced interviewers	8	5	0	61.5
Audio record each interview in total	7	6	0	53.8
Use Computer-Assisted Recorded Interviewing tools to record parts of interviews	5	7	1	41.7
Verify interview completion with respondents using in person visits	3	7	3	30.0
Verify interview completion with respondents using telephone calls	11	2	0	84.6
Verify interview completion with respondents using mailed postcards	2	10	1	16.7

Post-Data Collection Policies and Practices (may also be used during data collection)

Does your organization:	Yes	No	N/A	Valid %
Examine paradata related to the length of the questionnaire overall at the interviewer level with the goal of identifying unusual patterns	12	1	0	92.3
Examine paradata related to the length of individual sections in the questionnaire at the interviewer level with the goal of identifying unusual patterns	5	8	0	38.5
Examine other types of paradata (e.g., backups, answer changes) at the interviewer level with the goal of identifying unusual patterns	9	4	0	69.2
Examine paradata related to the length of individual questions in the questionnaire for individual respondents with the goal of identifying unusual patterns	11	2	0	84.6
Examine other types of paradata (e.g., backups, answer changes) for individual respondents with the goal of identifying unusual patterns	4	9	0	30.8
Examine other types of paradata (e.g., backups, answer changes) for individual respondents with the goal of identifying unusual patterns	7	6	0	53.8
Examine interviewer-level patterns of consent for pieces of a study in addition to the questionnaire (e.g., biomarkers, linkage to administrative records)	6	2	5	75.0
Examine interviewer-level patterns of results for biomarker data	4	2	7	66.7
Evaluate signatures on respondent incentive forms for authenticity	4	5	4	44.4

Other Organization Level Policies and Practices

Does your organization:	Yes	No	N/A	Valid %
Provide time off for interviewers and supervisors for family care	10	2	0	83.3
Provide time off for other research staff for family care	13	0	0	100.0
Provide paid sick leave time off for interviewers and supervisors	6	6	1	50.0
Provide paid sick leave time off for other research staff	12	1	0	92.3
Tie the promotion opportunities for research staff to their publication rate in peer-reviewed journals	1	10	2	9.1
Hire full time interviewers	6	7	0	46.2
Pay interviewers per completed interview	1	12	0	7.8
Pay interviewers per hour	12	1	0	92.3
Pay interviewers bonuses for completes	1	12	0	7.8
Pay interviewers bonuses for refusal conversion	2	11	0	15.4

APPENDIX D: Charge of the Task Force

Background

Within the last few years, there has been a spike in concerns about data falsification, especially interviewer falsification of survey data. In addition, new methods (such as computer-assisted recording of interviewers and tracking of interviewers via GPS) for detecting falsification have become available. Many of these new approaches involve statistical methods for detecting seemingly suspicious patterns of data across interviews.

Scope and Goals

The goal of this Task Force would be to produce a white paper on data falsification. The paper would review empirical efforts to assess the level of the problem, present brief case studies of highly publicized or highly damaging examples of data falsification (e.g., the Diederik Stapel case), examine and evaluate the various methods currently used for detecting falsification of survey data, and make recommendations regarding best practices for detecting falsification, including traditional ongoing monitoring efforts and more recent methods involving data analytic methods.

Composition of the Task Force and Terms

There would be two co-chairs of the Task Force, one representing AAPOR and the other representing ASA. The remaining 8 to 10 members of the committee would be divided equally among those appointed by the two organizations.

The Task Force member would serve two-year terms, which could be extended by the AAPOR Council (for the AAPOR appointees) or ASA's Board of Directors (for the ASA appointees), if more time were needed to complete the white paper.

The AAPOR co-chair would report periodically to AAPOR's Council on the Task Force's progress. The ASA co-chair would have similar reporting responsibilities to the ASA Board of Directors.