

How Worried Should We Be? The Implications of Fabricated Survey Data for Political Science

Oscar Castorena¹, Mollie J. Cohen², Noam Lupu¹ , and Elizabeth J. Zechmeister¹

¹Department of Political Science, Vanderbilt University, USA

²Department of Political Science, University of Georgia, USA

All correspondence concerning this article should be addressed to Noam Lupu, Vanderbilt University, PMB 0505, 230 Appleton Place, Nashville TN 37203-5723, USA. Email: noam.lupu@vanderbilt.edu

Abstract

Surveys are ubiquitous in the study of politics, making enumerator fabrication a critical issue. A prevailing view is that faked interviews affect inferences drawn from compromised datasets. Researchers have generated theories about how fabrication might affect inferences. Yet, speculation has outpaced systematic testing. We leverage a rare dataset to address this gap: a national face-to-face survey in Venezuela in which a uniquely high volume of falsified interviews was detected, canceled, and replaced. Comparing the verified and fraudulent datasets, we find that descriptive inference is sometimes affected, but correlational results hold, even in a dataset with an unusually high-fabrication rate. Enumerators largely fabricate plausible data. Though still egregious, enumerator fabrication may not constitute a grave threat to political science research.

Surveys are ubiquitous in the study of politics, making survey errors a critical issue for political science (Heath, Fisher, & Smith, 2005; Lupu & Michelitch, 2018). Among the most egregious errors is wholesale fabrication. Recent studies have focused on fabrication by either respondents who provide bogus responses or office employees who duplicate interview data (Kuriakose & Robbins, 2016; Pew Research Center, 2020). A more classic concern is the fabrication of entire interviews by fieldworkers in enumerator-administered surveys (e.g., Crespi, 1945; Gomila, Littman, Blair, & Levy Paluck, 2017). This issue is of paramount importance for political science; in research on many parts of the world and especially developing contexts, the most prominent surveys used in the discipline are administered in person by enumerators operating at a distance from a supervising team (Lupu & Michelitch, 2018). Yet, we know little about how the existence of enumerator-fabricated interviews might affect the discipline's conclusions, in large part because we know very little about wholesale fabrication by interviewers in the kinds of surveys analyzed by political scientists.

There is no doubt that interviewers sometimes fabricate data. In our own work, we have detected enumerators at home interviewing friends, discussing ways to avoid location monitoring, and posing as respondents by modulating voices on audio files recorded for quality control. Similar anecdotes have long-raised concerns about the quality of academic survey data (Menold, Winker, Storfinger, & Kemper, 2013). Some scholars have argued that even small numbers of fabricated interviews can generate important biases (e.g., DeMatteis et al., 2020; Gomila, Littman, Blair, & Levy Paluck, 2017). In addition, awareness of fabrication in surveys could affect public perceptions of scientific integrity (Johnson, 2018). For

example, in a recent public debate on this topic, one shocking headline read, “Many surveys, about one in five, may contain fraudulent data” (Bohannon, 2016). While the underlying study was more nuanced, the message to the public was alarming.

The research community has addressed fake data through revisions to best practices and innovations in fraud detection (AAPOR, 2003; Bredl, Storfinger, & Menold, 2013; Cohen & Warner, 2021; Crespi, 1945; Gomila, Littman, Blair, & Levy Paluck, 2017; Montalvo, Seligson, & Zechmeister, 2018; Robbins, 2018; Schraepfer & Wagner, 2003; Slomczynski, Powalko, & Krauze, 2017).¹ When faked interviews are discovered after a study is closed, they cannot be replaced with the intended authentic interview. But most approaches to detecting fake data are not well-suited to finding wholesale fabrication while surveys are in the field (DeMatteis et al., 2020; Finn & Ranchhod, 2017; Judge & Schechter, 2009; Kuriakose & Robbins, 2016; Schäfer, Schrapler, Muller, & Wagner, 2004). And when fraudulent data are identified, they are usually removed from published datasets. What, then, does enumerator fabrication imply for research on politics?

A prevailing view is that faked interviews affect the inferences drawn from compromised datasets. Although it appears that only a small number of interviews—perhaps less than 2–5%—are faked in typical large-scale surveys

¹ It is also important to decrease enumerators' incentives to cheat. For example, when interviewers are paid by completed interview or punished for not meeting quotas; when survey instruments are very long, repetitive, or include sensitive questions; and when enumerators are asked to travel long distances or conduct interviews in insecure contexts, they may be more likely to engage in the wholesale fabrication of interviews (Crespi, 1945; Menold, Landrock, Winker, Pellner, & Kemper, 2018; Winker, 2016; Winker, Kruse, Menold, & Landrock, 2015).

(Bredl, Storfinger, & Menold, 2013; Cohen & Larrea, 2018; Menold, Winker, Storfinger, & Kemper, 2013; Schrapler & Wagner, 2003), some simulations suggest that even low levels of fabrication can bias inferences (Gomila, Littman, Blair, & Levy Paluck, 2017; Sarracino & Mikucka, 2017; Schrapler & Wagner, 2003).² Scholarship suggests two reasons to think that cheating produces biased data. First, if fabrication is motivated by incentives to complete interviews quickly, enumerators should favor approaches that allow them to speed through the questionnaire, drawing responses away from the true population mean (Gomila, Littman, Blair, & Levy Paluck, 2017). Second, cheating interviewers may be more likely to select middling responses on scale items to avoid drawing scrutiny for too many extreme answers (Bredl, Winker, & Kotschau, 2012; Menold, Winker, Storfinger, & Kemper, 2013; Porras & English, 2004). In either case, we would observe different means and lower variance in faked data (Gomila, Littman, Blair, & Levy Paluck, 2017; Kosyakova, Olbrich, Sakshaug, & Schwanhaeuser, 2019).

Yet, it is also plausible that fabricated data barely differ from real data. Cheating interviewers may know how authentic responses tend to look based on prior experience conducting similar surveys (Waller, 2013)—and use this knowledge to effectively mimic real data (Landrock, 2017; Menold, Winker, Storfinger, & Kemper, 2013; see also discussion in Blasius & Thiessen, 2021).

Assessing the effect of fabricated survey data on political science research is complicated by the absence of a counterfactual—the authentic data cheaters would have gathered from real respondents. As a result, most scholarship on interviewer fabrication relies on datasets with a relatively small number of faked interviews per sample (e.g., Schrapler & Wagner, 2003), comparisons between faked and real interviews conducted at different times and places (e.g., Kosyakova, Olbrich, Sakshaug, & Schwanhaeuser, 2019), simulations (e.g., Sarracino & Mikucka, 2017; DeMatteis et al., 2020), or data created by research assistants directed to fabricate responses (e.g., Landrock, 2017; Menold, Winker, Storfinger, & Kemper, 2013; Rosmansyah, Santoso, Bani Hardi, Putri, & Sutikno, 2019; but see Finn & Ranchhod, 2017). Researchers have theories about how large-scale fabrication might affect inferences, but speculation has outpaced systematic testing.

We leverage a rare opportunity to address this empirical gap. In overseeing a nationally representative, face-to-face survey in Venezuela in 2016/17, we detected an unusually high volume of falsified interviews and canceled and replaced them while fieldwork was in progress.

Most approaches to detecting fabricated data remove fraudulent cases after fieldwork is complete (e.g., Blasius and Thiessen, 2021; Kosyakova, Olbrich, Sakshaug, & Schwanhaeuser, 2019). But these approaches, while valuable, make it impractical or impossible to replace fabricated interviews with real ones. Our approach differs significantly because it allows us to detect fabrication in real time. While interviewers are in the field, we use dozens of automated and manual quality control checks to cancel and immediately

² Even studies examining interview falsification rarely report the prevalence of fake interviews in survey data. In other cases, it is not possible to estimate the prevalence of fraud based on available information. For example, while Turner, Gribble, Al-Tayyib, and Chromy, (2002) show that six enumerators falsified at least 49% of their total workload during the 1997–1998 Baltimore STD and Behavior Study, it is unclear what proportion of the total data collection these interviews represent. Of course, there are exceptions to this typical outcome (e.g., Bredl, Winker, & Kotschau, 2012).

replace fraudulent interviews. Following this approach, our dataset includes both the canceled and replaced interviews, which allows us to make direct comparisons between a validated dataset and the compromised one that would have resulted had the faked interviews not been replaced.

We find that descriptive inference is sometimes affected, but that correlational results hold in some typical applications, even in a dataset with an unusually high proportion of fabricated cases. Replication with a second dataset, from a similar study we fielded in Peru in 2017, yields similar results. Enumerators largely seem to fabricate plausible data, which tamps down on the likelihood that faked interviews severely threaten political science research. This matters because scholars have long relied on survey data to generate insights into public opinion and political behavior. The implication of our study is this: analyses of opinion datasets compromised by even a nontrivial degree of interviewer-generated fabrication can still generate valid conclusions.

The Venezuela Dataset

Venezuela experienced acute crises in 2016, including efforts to recall the president, civil unrest (including frequent protests and looting), and deteriorating economic conditions (McCarthy, 2017). This context of scarcity, unrest, and insecurity provided ample motivation for interviewers to skirt survey protocols.³

We fielded a national face-to-face survey from November 2016 to February 2017, using a reputable Venezuelan survey firm.⁴ The study used e-devices and specialized software for data collection supplemented by extensive quality control based on audio recordings, interviewer identity verification, timing features, and geographic coordinates, among other checks. The approach we used is similar to that reported by Cohen and Larrea (2018) and Montalvo, Seligson, and Zechmeister (2018).

With respect to audio recordings, our approach mirrors that used by Gomila, Littman, Blair, and Levy Paluck (2017): interviewers (and respondents) were informed that portions of the interview would be recorded, but not which questions were being observed in this way. During a two-day mandatory training, enumerators were informed of monitoring protocols, and were shown the quality control dashboard. Enumerators were also aware that other enumerators had been caught and removed from the project for falsifying interviews. In short, enumerators were both aware and had reason to believe that they were being monitored. Their location was monitored by building geofences into the software used for data collection, and violations were automatically flagged.

Suspicious interviews (whether because of audio, location, timing, or other flags) often were further investigated using additional information, including a front-facing picture that the device took in the course of the interview, providing additional insight into whether the interviewer was in the field or in a car, in a mall, or—in one case—in bed. The

³ Blasius and Thiessen (2021) note that enumerator fabrication is more likely in more corrupt contexts, and Winker (2016) notes that insecurity also drives fabrication. In 2017, Venezuela ranked among the worst in the Latin American region on numerous indicators of corruption perception and experience (Pring, 2017).

⁴ The survey firm was not required to disclose details about how payment to interviewers was administered. In our experience, the most common approach to compensating interviewers in Venezuela and other developing contexts is on a per-interview basis. This approach runs counter to best practice recommendations because it increases incentives to fabricate interviews or shirk in other ways that undermine data quality (AAPOR/WAPOR, 2021).

data-collection team audited all of the interviews at least once, while fieldwork was in progress. In just over one-third of the cases, a second individual re-audited interviews, to ensure that the auditors were themselves being audited. A large team of undergraduate research assistants worked on the project, reviewing the audio files on a daily basis.

Thus, in contrast to approaches that rely on post-fieldwork statistical analyses, we were able to identify, cancel, and replace interviews while enumerators were still in the field, with an average difference of 39 days between the fabricated and genuine interviews analyzed in this article. More than 650 interviews out of 1,500 were canceled and replaced due to quality concerns—a rate far higher than in any other survey of which we are aware.⁵

The study is based on an area probability sample that was stratified by four main geographical regions, size of municipality (small, medium, large), and urban and rural areas. The realized sample consists of 83 primary sampling units (PSUs; municipalities or parts of large municipalities) and 250 final sampling units (blocks). Selection processes were probabilistic down to the PSU, and systematic selection was used for blocks (northeast corner of the segment) and houses (one house skipped after each completed interview). Interviewers were instructed to complete six interviews per block. A frequency matching approach was used to select individuals within households, balancing on age cohorts and gender according to known population proportions. Interviewers could fill these age and gender quotas in any order; the balance of stipulated age and gender matches was monitored and enforced at the PSU level.⁶ If two or more people of the same gender and age group were present in the household at the moment of within-household selection, the survey was administered to the person who had most recently celebrated a birthday. Under this design, failure to achieve a particular interview is driven by availability of the targeted type (by gender and age) when interviewers are present and individuals are willing to participate.

Fabrication was not the only reason an interview was canceled and replaced. Some interviews were canceled for quality reasons—because an interviewer misread questions, for instance—but were likely real interviews. We used the quality control data and auditors' notes to identify 460 of the canceled interviews as fraudulent (see [Appendix B](#) for coding procedures). The remaining 190 canceled interviews are not analyzed here. To compare fraudulent and authentic cases, we paired each fraudulent case to an authentic interview from the final, published dataset with an exact match on gender, age group, and PSU.⁷ The result is a set of 420 fraudulent

⁵ For example, in the 2016/17 round of data collection for the AmericasBarometer project, on average only 2% of interviews were determined to be fabricated (Cohen & Larrea, 2018). There are of course exceptions that report higher rates (e.g., Bredl, Winker, & Kotschau, 2012).

⁶ Interviewers have access to a tab where they can track quota categories for their assigned PSU. The categories populate as soon as an interview is completed. If more than one interviewer is working in a PSU, each can see when quotas are filled.

⁷ We used the *cem* package in Stata to implement the exact matching. Results are robust to matching on education and income, although including these variables substantially reduces statistical power. See [Appendix B](#) for more details. This matching strategy is similar to that employed by Menold, Winker, Storfinger, and Kemper (2013) and Rosmansyah, Santoso, Bani Hardi, Putri, and Sutikno (2019). Those studies match interviews falsified in the lab and during enumerator training, respectively, to interviews later conducted in the field. Our falsified interviews may thus follow different patterns and respond to different incentives compared to such “fake-fake” data. [Table A4](#) shows the geographic distribution of fraudulent interviews and reveals that they were somewhat more likely in the Capital region than elsewhere in the country.

Table 1. Item-Level Effects of Fabricated Data

Comparison	Fake versus Matched	Compromised versus Clean
Difference in means	11.5–48.7%	0.0–13.3%
Average magnitude (in SD)	0.20–0.31	0.08–0.11
Difference in variances	8.9–46.0%	0.9–4.4%
Item nonresponse	0.0–20.4%	0.0–0.9%

Note. Values result from tests of 113 items, comparing the fabricated interviews and the matched real data ($N = 420$) as well as the compromised data and the clean data ($N = 1,489$). In each case, we generate results using no adjustment to the standard errors as well as the Bonferroni correction, Hochberg's step-up procedure, and Holm's step-down procedure—and we report the range of values. We use a cutoff of $p < .10$ for statistical significance. Full results in [Tables A2 and A5](#).

interviews that are precisely matched to 420 authentic interviews, allowing us to make direct comparisons between two scenarios: one in which the fabrication had been allowed to compromise the dataset and the counterfactual, a dataset consisting only of validated interviews.⁸

Does Fraud Make a Difference?

Do falsified interviews differ from valid interviews? We focus on the means and distributions of questions with ordinal response scales, which constitute 86% of the attitudinal items on the survey.⁹ For each of these 113 items, we examine differences in means, variances, and item nonresponse rates. We generate results using four different estimates of standard errors to account for multiple testing: using no adjustment, the Bonferroni correction, Hochberg's step-up procedure, and Holm's step-down procedure. To minimize the potential for false negatives, we use a generous cutoff of $p < .10$ for statistical significance (the full results are provided in [Supplementary Tables A2 and A5](#)).

[Table 1](#) reports the ranges of the resulting values for these 113 items. The fake versus matched column reports on tests that compare only the fabricated interviews to their matched authentic interviews, while the compromised versus clean column reports on tests that compare the national survey containing the fraudulent interviews (and not their replacements) and the full national survey with only valid interviews. When comparing the fabricated interviews to their genuine counterparts, between 13 (11.5%) and 55 (48.7%) items show significant mean differences between the fraudulent and clean interviews, depending on which standard errors we estimate. On average, the magnitude of these significant differences is between 0.20 and 0.31 standard deviations. With respect to the variance of responses, for 82% of the items we examine,

⁸ We focus on interviews identified to have been entirely fabricated. However, it is possible that some interviewers fabricated portions of interviews that were accepted and published (see [De Haas & Winker \(2016\)](#) for an analysis of partial interview fabrication). We assessed the validated dataset for this possibility by testing for straightlining in a portion of the survey where we would expect to observe shirking behavior: a series of 18 questions with seven-point response scales. We found that across the validated dataset, only three observations had nondifferentiated response patterns (i.e., they had the same response for each item). We note that this minimal straightlining behavior could be the result of either enumerator fraud or respondent satisficing ([Krosnick, 1991](#)).

⁹ The remaining survey items are nearly all questions that use categorical response options. Because there are very few continuous variables in the dataset, we do not use Benford's Law to test for fraud here (see [Schrapler & Wagner \(2003\)](#) for an example).

the standard deviation of fraudulent responses is smaller than the standard deviation of the matched clean responses. Variance ratio tests assessing differences in the standard deviations indicate significant differences between 10 (8.9%) and 52 variables (46.0%). We find between 0 and 23 items with significant differences in item nonresponse rates, depending on which standard errors we compute.¹⁰

On these metrics, enumerator-faked data differ from real data, sometimes in a substantial portion of questions, depending on how we estimate our standard errors. The differences are generally fairly small in magnitude, and of course, they are further attenuated when comparing the full clean dataset to a compromised dataset that substitutes genuine interviews from the final validated dataset with their matched fraudulent interviews. As shown in the rightmost column in [Table 1](#), the proportion of items in the full dataset with significant differences is small, regardless of how we estimate our standard errors.

Does Fraud Affect Inferences?

Most research draws inferences from the patterns observed in data. One approach is to compare means over time, in order to report on change and continuity in attitudes. We examined differences across time for the 49 items that were also included in a similar survey conducted in Venezuela in 2014. Would we have reached different conclusions had we not identified and replaced the fake data? The answer is yes, though only 14.3% of the time (7 of 49 variables) and not because the cross-time differences are statistically distinct from each other, but because results fell just inside or just outside the 95% confidence threshold (see [Supplementary Figure A1](#)).

The 2016/17 Venezuela dataset also allows us to assess how a substantial amount of fake data affects standard political science regression models. We compare the results from identical regression analyses that use the final, clean dataset and the compromised dataset that replaces genuine responses with the matched fraudulent responses. We selected two models common to comparative political behavior research. The first predicts support for democracy, an attitude of central concern in public opinion scholarship (e.g., [Evans & Whitefield, 1995](#); [Mishler & Rose, 1996](#)). The second model predicts political tolerance, following seminal work by [Gibson \(1992\)](#); among others, see [Golebiowska, 1999](#); [Shamir & Sullivan, 1983](#)). The support for democracy model includes independent variables whose means significantly differ between matched genuine and fabricated data, while the political tolerance model uses a dependent variable whose mean differed significantly (see [Tables 1](#), [Supplementary Tables A1 and A2](#)).

[Figure 1](#) plots the regression coefficients. The dependent variable in the left-hand panel is based on a question that asks the level of agreement with the statement, Democracy may have problems, but it is better than any other form of government. Following previous research (e.g., [Seligson, 2007](#); [Singer, 2018](#)), our model includes measures of respect for political institutions, presidential approval, past vote choice, left-right self-placement, perceived neighborhood insecurity, crime victimization, trust in police, evaluation of the national economy, evaluation of one's personal economic situation, gender, age, education, and skin tone (all independent

¹⁰ Our results with corrected standard errors contrast with [Gomila, Littman, Blair, and Levy Paluck \(2017\)](#), who find evidence, in a survey in Nigeria, that interviews that failed their quality control checks had higher rates of item nonresponse.

variables are scaled from zero to one; for wording and coding, see [Supplementary Table A3](#)). The results are broadly similar across the clean and compromised datasets. The coefficients for three variables change signs across the analyses (nonvoter in the 2013 elections, national economic evaluation, and skin tone), but none of these coefficients are significantly different from zero. Overall, both analyses support similar conclusions that are consistent with prior studies of support for democracy in the region.

For the political tolerance model (right-hand panel of [Figure 1](#)), the dependent variable is measured using an item that asks, "How strongly do you approve or disapprove that [regime critics] be allowed to conduct peaceful demonstrations in order to express their views?" Following [Duch and Gibson \(1992\)](#) and [Gibson \(1992\)](#), our models include measures of perception of threat, political efficacy, level of education, age, and gender. Again, the results are broadly similar across the clean and compromised datasets. The coefficient sign is different for only one variable (age), and that estimated coefficient is essentially zero. Overall, both datasets yield results that are in line with prior research: threat, lower efficacy, and lower education are associated with lower tolerance.

There is also similarity across coefficient parameters estimated using the clean versus compromised datasets. Chow tests of whether a coefficient estimated from the clean dataset is equal to the corresponding coefficient estimated from the compromised dataset yield no statistically significant differences (see [Supplementary Tables A7 and A9](#)).

There are instances where slight differences could affect conclusions. For example, in the support for democracy analysis, the coefficient for Center is significant in the clean dataset, but not significant in the compromised dataset. And the coefficient for education in the tolerance analysis using the compromised dataset is outside the standard cutoff, while the corresponding coefficient using the clean dataset is within it.¹¹ From one perspective, the mere existence of some differences is reason for concern. From another perspective, this general picture is reassuring. Even unusually large amounts of fraud make only a small difference in the inferences we draw. Notably, we find that inferences that rely on small effect sizes or estimates that are close to the threshold for statistical significance—results we should already treat with some skepticism—are most likely to be affected by compromised data.

How Do Interviewers Fake Data?

Concerns about the ill effects of fake survey data derive from the expectation that enumerators fabricate interviews in ways that substantially bias the data. Our results suggest that, in fact, interviewers create fake data that has a fairly high degree of internal consistency, such that the correlations between variables are realistic. Enumerators who fabricate interviews appear to mimic reality.

How do faking interviewers know what reality to mimic? The data suggest that they employ a mixed strategy, which allows them to learn about actual public opinion from real interviews before falsifying interviews. The Venezuela survey employed 79 interviewers; of these, 46 fabricated interviews.

¹¹ If we used a threshold of $p < .1$ for statistical significance, we would conclude that Vote for Maduro and Trust in National Police are significant predictors in the clean dataset, while the result would be null (a false negative) in the compromised dataset (see [Table A6](#)). We report analyses using a $p < .05$ threshold here since our aim is to understand how fabricated data may affect inferences in typical public opinion analyses.

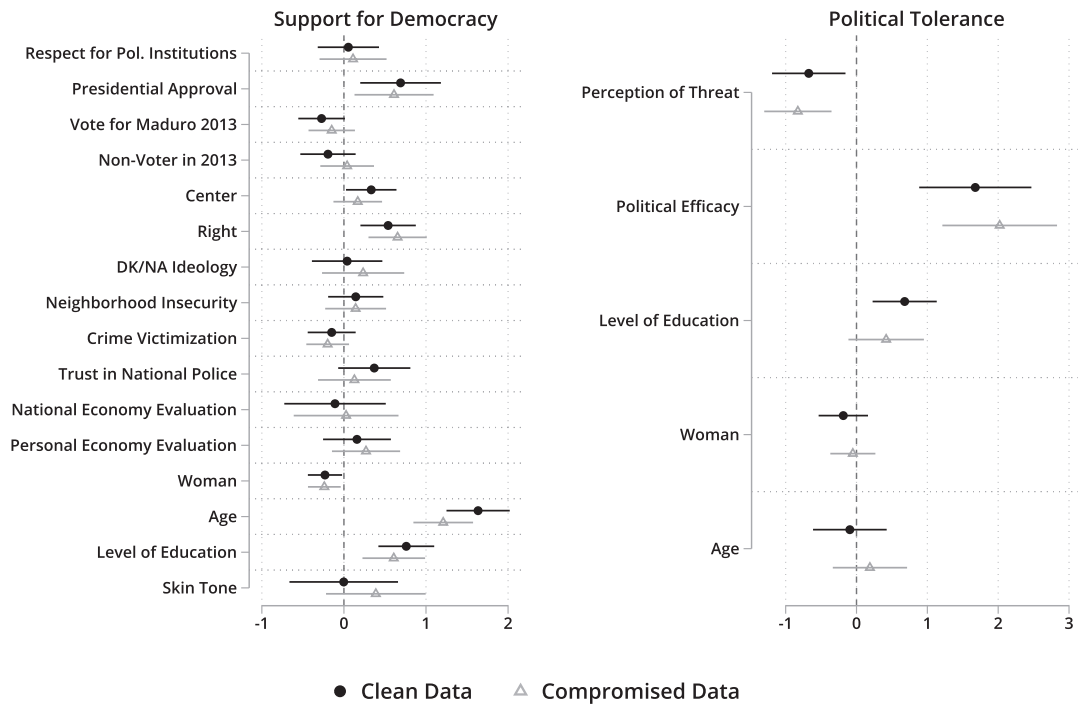


Figure 1. Comparing Regression Results. Values represent coefficients estimated from OLS models using both the clean and compromised datasets with 95% confidence intervals. Standard errors account for sampling design effects.

Prior to their termination, these enumerators averaged 10 fake interviews but also 21.5 real interviews. Most fabricators did so only after conducting real interviews: 50% of enumerators who eventually recorded a falsified interview had conducted at least five real interviews before their first fraudulent interview (only 25% of all cheaters falsified their first recorded interview). That is, most enumerators learned about response patterns prior to fabricating interviews.

Given such a strategy, we might expect that the faked observations replace the kinds of interviews that are hardest to obtain (DeMatteis et al., 2020; Kosyakova, Skopek, & Eckman, 2014; Winker, 2016). The sample design for the survey used gender- and age-based frequency matching for respondent selection within households. This allows us to test the assumption that interviewers are motivated to replace hard-to-access respondents with fake interviews. In many surveys, a key challenge is recruiting working-age men, since they are frequently not home and less likely to participate (Silver, McRoy, Devlin, & Moynihan, 2019). Recall that, given the sample design, interviewers were instructed to achieve a certain balance with respect to gender and age, and that balance was monitored and enforced at the PSU level. If interviewers are faking interviews when it becomes especially difficult to conduct real ones, and a certain subgroup—e.g., working-age men—is disproportionately harder to find and recruit, we would expect interviewers to have a greater incentive to fabricate those interviews. As a result, the fabricated data should be disproportionately male and working-age.

This is indeed a pattern we observe. Table 2 compares the demographic composition of fabricated interviews with that of the full, clean data. The fraudulent interviews contain slightly more young and middle-aged men and fewer older men and women of all age-groups. In sum, interviewers do appear to fabricate data partly in order to complete difficult fieldwork assignments.

Finding that cheating interviewers create plausible fake interviews runs counter to what some might theorize. Research suggests that satisficing leads individuals to select responses in the middle of a scale or take the shortest path through the instrument (e.g., see discussions in Bredl, Storfinger, & Menold, 2013; Gomila, Littman, Blair, & Levy Paluck, 2017; Menold, Winker, Storfinger, & Kemper, 2013; Schrapler & Wagner, 2003). If the enumerators used these approaches—or mere random selection—we would find far more differences. To demonstrate, we compare the observed differences in the fraudulent data to what would have been produced under other strategies. First, we generated a version of fraudulent data by programming a random selection of responses. Next, we hired 10 students and incentivized them to speed through the questionnaire as quickly as possible with devices similar to those used in the fieldwork, each 10 times. By offering an incentive to the fastest speeder, we aimed to generate data that identified the fastest path through the electronic questionnaire; while skips can be observed directly, the ways that the question wording and its response options appear on the screen varies across questions, making it impossible to infer the fastest path without putting it to an empirical test. Finally, we created a version of the fraudulent data based on selecting middling responses (these were normally distributed with a mean at the center of a scale and standard deviation of one tenth of its range).

Table 3 presents the comparison between each of these simulations and the matched authentic interviews. Again, we report ranges for each value based on the four methods of estimating standard errors. In every case, we find far higher mean differences, average difference magnitudes, variance ratios, and item nonresponse than we do in the observed falsified interviews. In other words, interviewers seem not to hew very closely to any of these alternative hypothesized strategies when they fabricate interviews. (We do find some evidence

Table 2. Demographic Composition of Clean and Fake Data

Age group (in years)	Male			Female		
	Fraud (%)	Clean (%)	Diff.	Fraud (%)	Clean (%)	Diff.
18–25	12.6	11.0	+1.6	10.0	10.1	–0.1
26–35	11.7	11.2	+0.5	10.9	11.0	–0.1
36–45	11.7	9.8	+1.9	6.7	10.7	–4.0
46–55	13.6	8.5	+5.0	7.9	8.4	–0.5
56–65	5.0	5.9	–0.9	5.2	6.0	–0.8
66+	1.9	3.9	–2.0	2.9	3.5	–0.6

Note. Values compare the fraudulent data to the full clean dataset.

that faking interviewers shy away slightly from extreme response options; see [Supplementary Figures A2 and A3.](#))

A final way to examine the fabrication production process is to consider the structural similarities between falsified and real interviews. If interviewers are mimicking true responses, fake data should look similar to real data—and may even underestimate population heterogeneity. We can see this by running [Kuriakose and Robbins' \(2016\)](#) *percentmatch* program, which assigns a score to each interview based on the extent to which responses match those of the most similar interview. This program is designed to detect near duplicates, generated by cutting-and-pasting survey data rows to generate additional interviews. When it occurs, this type of fabrication is typically generated in an office because it requires access to the raw dataset. For each pair of interviews in the dataset, *percentmatch* calculates the percentage of observed answers that are identical. [Kuriakose and Robbins \(2016\)](#) argue that interviews with a match score of 85% or above (i.e., those with identical answers to 85% or more questions with at least one other interview) are suspicious.

[Figure 2](#) summarizes the *percentmatch* scores for the fraudulent interviews and the clean interviews with which they are matched (as in [Table 2](#)). Overall, the pattern across the two is quite similar. The fraudulent cases exhibit greater similarity in their responses to other interviews. This evidence is consistent with a process by which interviewers glean information from the interviews that they conduct and/or their general knowledge of the population and draw on this when they fabricate a set of plausible, but slightly more internally consistent, responses.¹²

It is important to note that while we see some evidence that interviewers gather information by conducting some real interviews within our study before faking data, we are unable to test more systematically the extent to which that process is crucial to the results. This is because we cannot observe whether interviewers entered the study already in possession of reasonable priors about the general population. For example, if most interviewers were experienced in similar surveys, their tendency to conduct valid interviews prior to real ones could reflect efforts to evade suspicion rather than efforts to discern opinion patterns. Our core thesis is that when enumerators fake interviews, they do so with some knowledge

¹² [Figure 2](#) also shows that *percentmatch* does not detect enumerator-generated fabrication: all of the scores are below the 0.85 threshold for likely fabrication. [Figure A4](#) compares the distribution of *percentmatch* scores within the fraudulent data to that within the entire dataset of clean interviews.

Table 3. Item-Level Effects of Fabricated Data

Comparison	Faked	Random	Speeding	Middling responses
Difference in means	11.5–48.7%	68.0–81.4%	75.2–86.7%	71.7–85.8%
Average magnitude (in SD)	0.20–0.31	0.65–0.75	0.73–0.82	0.63–0.72
Difference in variances	8.9–46.0%	26.6–66.4%	77.9–92.9%	72.6–98.2%
Item non-response	0.0–20.4%	–	50.4–78.8%	–

Note. Values result from tests of 113 items. In each case, we generate results using no adjustment to the standard errors as well as the Bonferroni correction, Hochberg's step-up procedure and Holm's step-down procedure—and we report the range of values. We use a baseline cutoff of $p < .10$ for statistical significance. The full results are in [Table A10](#).

about the distribution and consistencies in real interviews—and they try to mimic real data. Our conclusion is that this informed fabrication process largely undergirds the patterns and relationships found in genuine data.¹³

How Worried Should We Be?

There is no question that interview fabrication in survey data is egregious. Our analyses affirm what other studies have found: that fabrication can result in some differences in final datasets. Still, the consequences for inference may be less severe than many studies suggest. Widespread fabrication is rare. But our analyses reveal that, even in unusual instances where fabrication is extensive, its effects on the kinds of analyses that political scientists typically conduct may be minimal. Fabrication affects descriptive statistics for a small proportion of variables. Correlations, the basis of most political science applications of survey data, are remarkably similar across both clean and fake data. To the extent that there are differences, they emerge for coefficients that are borderline significant and ought, already, to be treated with caution.

The consequences of fabricated data are lower to the degree that interviewers use their knowledge of real data when fabricating interviews. In contrast to expectations that cheating enumerators might choose responses at random, to maximize speed, or toward the middle of the response range, we find interviewers tend to mimic real data, basing their false responses on real interviews they conduct prior to fabricating others.¹⁴ As a result, even a large proportion of fake interviews has limited effects on researchers' inferences.¹⁵ It is implausible to assert that datasets used in political science research, in particular older datasets, are free from interviewer-generated fabrication. While certain datasets may be compromised, it does not follow that conclusions based on them are also compromised—rather, to the extent that

¹³ While this section has considered *how* interviewers fake data, another important question is *when* they do so. In [Appendix E](#), we show that fabrication sharply increased just before a holiday, when fieldwork was initially scheduled to end, and that it was more common later in the day. Given the high prevalence of fabrication in this study and the economic and security challenges of the context, we surmise that fabrication is more likely when interviewers face more challenging field conditions. Pay structure likely also matters (see footnotes 1 and 4). Future research on this topic should work to advance knowledge regarding when fabrication is most likely to occur and, as well, how that timing intersects with the nature of the fabricated data. We do not have sufficient information about interviewers' characteristics or level of experience to examine *who* fakes data, although we hope that future work takes up these questions.

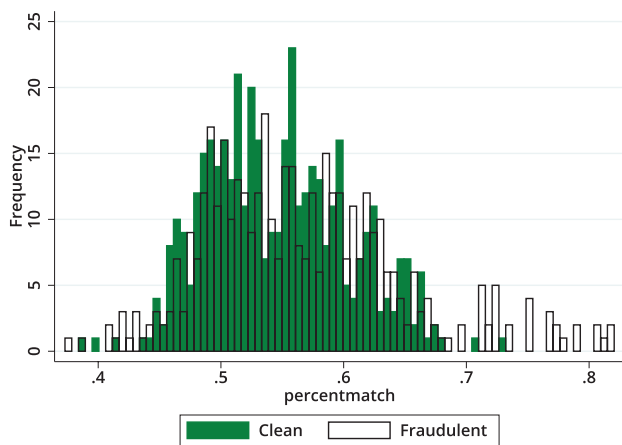


Figure 2. Similarity Among Fraudulent and Real Interviews. Histogram of percentmatch scores for matched clean and fraudulent interviews. A value of one indicates complete duplication between an interview and its most similar interview.

the data-generating process for the faked data mirrors what we describe here, the community can continue to draw reliable insights from such datasets.

We rely, however, on data from just one study, so readers might ask how generalizable our results are. The political and socioeconomic context in Venezuela in 2016–17 was troubled, creating incentives for enumerators to shirk. But it was not unlike other developing contexts in which researchers regularly field face-to-face surveys. Our monitoring protocols are more extensive than those employed by most researchers, but nearly all face-to-face survey researchers employ some type of monitoring (Lupu & Michelitch, 2018), dating back to Crespi's (1945) important contribution to enumerator cheating.

We can assess the question of generalizability empirically by using data from another case in which we identified, canceled, and replaced a sufficient number of interviews to replicate the analyses in this paper. In 2017, we employed the same training and quality control protocols as we conducted a national survey of Peru. The sampling approach was identical to the Venezuela study, including the selection of individuals within the household and the enforcement of gender and age quotas at the PSU level. We found—and replaced in real time—116 likely fabricated interviews (4.4% of the sample)—a rate that is higher than the estimated level of fabrication in the average scientific public opinion survey (Bredl, Storfinger, & Menold, 2013; Cohen & Larrea, 2018). We matched these data to authentic interviews and the results affirm our conclusions (see Supplementary Table A12 and Figure A11). The average difference in means is of a similar magnitude, though fewer of these differences are significant given the smaller number of faked interviews. Of course, it remains possible that our supervision protocols—and the fact that interviewers were told about them in advance—affected whether and how interviewers engaged in fraud.

Still, these findings are reassuring for scholarship that relies on face-to-face surveys, particularly those that are unable to

use extensive protocols for fraud detection or replace fraudulent interviews while fieldwork is ongoing. We do not advocate being cavalier about fabricated data. Researchers should take all possible measures to avoid and detect fraudulent data, and fraudulent interviews should whenever possible be removed and, ideally, replaced with real ones.

While interviewer-generated fabrication may be a predominant cause of fake data, there is heterogeneity in the motives and methods that produce such data. Data fabricated in an office by (near) duplication of observations can decrease variance and, thus, confidence intervals, increasing the potential for Type I (false positive) errors (Kuriakose & Robbins, 2016). Our conclusions do not travel to these situations, or to others such as data faked by firms seeking to please clients with particular results. It is critical, then, for survey researchers to be transparent about quality control processes and other information relevant to surmising the likelihood that different types of fabrication may be present. Ensuring the accuracy and quality of survey data is paramount.

Our results do highlight that data fabricated by interviewers could change inferences about fine-grained comparisons or small effects near standard statistical thresholds. But our findings also suggest that skepticism about the reliability of interviewer-administered survey data—especially from international surveys—may be exaggerated when it comes to the implications for political science. Scientific research should always strive for accuracy, but our research reveals that we can still learn a great deal from survey data even when some of it is fake.

Funding

Funding for data collection came from the United States Agency for International Development (Agreement AID-0AA-A-16-00021) and the Universidad Católica Andrés Bello. Neither funder had any role in this study's design, analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

For their comments and advice, the authors are grateful to Michael Robbins and audiences at Fundação Getúlio Vargas, GESIS, Universidad de Los Andes, and Vanderbilt. The authors also thank LAPOP Lab staff and research assistants along with their partners in Venezuela for their extraordinary efforts in collecting these data and Facundo Salles Kobilanski for excellent research assistance. The authors are grateful to the National Endowment for Democracy's Center for International Media Assistance and other supporters of the AmericasBarometer. They presented previous versions of this article at the 2019 conference of the Midwest Association for Public Opinion Research, the 2020 BigSurv conference, the 2020 meeting for Latin American Political Methods, and the 2022 MapleMeth conference.

Author Note

Oscar Castorena is a senior statistician at LAPOP Lab at Vanderbilt University.

Mollie Cohen is an assistant professor of Comparative Politics at the University of Georgia.

¹⁴ This implies that the effects of fabricated data may be different for studies in which interviewers have very little knowledge about or prior interaction with the target population—as when international rather than local enumerators are used.

¹⁵ Our findings can only speak to cases of wholesale interview fabrication; it is possible that partial interview fabrication might have different effects on study results.

Noam Lupu is an associate professor of Political Science and associate director of LAPOP Lab at Vanderbilt University.

Elizabeth J. Zechmeister is Cornelius Vanderbilt professor of Political Science and director of LAPOP Lab at Vanderbilt University.

Conflicts of Interest

The authors declare no conflicts of interest.

Supplementary Data

Supplementary data are available at the *International Journal of Public Opinion Research* online.

Notes

References

- AAPOR/WAPOR. (2021). *Task force report on quality in comparative surveys*. Retrieved from https://wapor.org/wp-content/uploads/AAPOR-WAPOR-Task-Force-Report-on-Quality-in-Comparative-Surveys_Full-Report.pdf
- American Association for Public Opinion Research (AAPOR). (2003). *Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects*. Retrieved from http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf
- Blasius, J., & Thiessen, V. (2021). Perceived corruption, trust, and interviewer behavior in 26 European countries. *Sociological Methods & Research*, 50(2), 740–777. doi:10.1177/0049124118782554
- Bohannon, J. (2016, February 24). Many surveys, about one in five, may contain fraudulent data. *Science*.
- Bredl, S., Storfinger, N., & Menold, N. (2013). A literature review of methods to detect fabricated survey data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention* (pp. 3–24). Frankfurt, Germany: PL Academic Research.
- Bredl, S., Winker, P., & Kötschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology*, 38(1), 1–10.
- Cohen, M. J., & Larrea, S. (2018). *Assessing and improving interview quality in the 2016/17 AmericasBarometer (LAPOP Lab Methodological Note IMN002)*. LAPOP Lab.
- Cohen, M. J., & Warner, Z. (2021). How to get better survey data more efficiently. *Political Analysis*, 29(2), 121–138. doi: 10.1017/pan.2020.20
- Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly*, 9(4), 431–445. doi:10.1086/265760
- De Haas, S., & Winker, P. (2016). Detecting fraudulent interviewers by improved clustering methods—the case of falsifications of answers to parts of a questionnaire. *Journal of Official Statistics*, 32(3), 643–660. doi:10.1515/jos-2016-0033
- DeMatteis, J. M., Young, L. J., Dahlhamer, J., Langley, R. E., Murphy, J., Olson, K., ... Sharma, S. (2020). *Task force report: Falsification in surveys*. AAPOR Council and the Executive Committee of the American Statistical Association. Retrieved from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Data_Falsification_Task_Force_Report-updated.pdf
- Duch, R. M., & Gibson, J. L. (1992). 'Putting up with' fascists in Western Europe: A comparative, cross-level analysis of political tolerance. *Western Political Quarterly*, 45(1), 237–273. doi:10.2307/448773
- Evans, G., & Whitefield, S. (1995). The politics and economics of democratic commitment: Support for democracy in transition societies. *British Journal of Political Science*, 25(4), 485–514. doi:10.1017/S0007123400007328
- Finn, A., & Ranchhod, V. (2017). Genuine fakes: The prevalence and implications of data fabrication in a large South African survey. *World Bank Economic Review*, 31(1), 129–157. doi:10.1093/wber/lhw054
- Gibson, J. L. (1992). Alternative measures of political tolerance: Must tolerance be 'least-liked?' *American Journal of Political Science*, 36(2), 560–577. doi:10.2307/2111491
- Golebiowska, E. A. (1999). Gender gap in political tolerance. *Political Behavior*, 21(1), 43–66. doi:10.1023/A:1023396429500
- Gomila, R., Littman, R., Blair, G., & Levy Paluck, E. (2017). The audio check: A method for improving data quality and detecting data fabrication. *Social Psychological and Personality Science*, 8(4), 424–433. doi:10.1177/1948550617691101
- Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science*, 8, 297–333. doi:10.1146/annurev.polisci.8.090203.103000
- Johnson, T. P. (2018). Presidential address: Legitimacy, wicked problems, and public opinion research. *Public Opinion Quarterly*, 82(3), 614–621. doi:10.1093/poq/nfy029
- Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources*, 44(1), 1–24. doi:10.3368/jhr.44.1.1
- Kosyakova, Y., Olbrich, L., Sakshaug, J., & Schwanhaeuser, S. (2019). *Identification of interviewer falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany*. Institute for Employment Research FDS Methods Report. Retrieved from http://doku.iab.de/fdz/reporte/2019/MR_02-19_EN_.pdf
- Kosyakova, Y., Skopek, J., & Eckman, S. (2014). 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
- Krosnick, J. A. (1991). Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236. doi:10.1002/acp.2350050305
- Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, 32(3), 283–291. doi:10.3233/SJI-160978
- Landrock, U. (2017). Explaining political participation: A comparison of real and falsified survey data. *Statistical Journal of the IAOS*, 33(2), 447–458. doi:10.3233/SJI-160270
- Lupu, N., & Michelitch, K. (2018). Advances in survey methods for the developing world. *Annual Review of Political Science*, 21, 195–214. doi:10.1146/annurev-polisci-052115-021432
- McCarthy, M. M. (2017). Venezuela's manmade disaster. *Current History*, 116(787), 61–67. doi: 10.1525/curh.2017.116.787.61
- Menold, N., Landrock, U., Winker, P., Pellner, N., & Kemper, C. J. (2018). The impact of payment and respondents' participation on interviewers' accuracy in face-to-face surveys: Investigations from a field experiment. *Field Methods*, 30(4), 295–311. doi:10.1177/1525822X18783977
- Menold, N., Winker, P., Storfinger, N., & Kemper, C. J. (2013). A method for ex-post identification of falsifications in survey data. In P. Winker, N. Menold, & R. Porst (Eds.), *Interviewers' deviations in surveys: Impact, reasons, detection and prevention* (pp. 25–48). Frankfurt, Germany: PL Academic Research.
- Mishler, W., & Rose, R. (1996). Trajectories of fear and hope: Support for democracy in post-communist Europe. *Comparative Political Studies*, 28(4), 553–581. doi:10.1177/0010414096028004003
- Montalvo, J. D., Seligson, M. A., & Zechmeister, E. J. (2018). Improving adherence to area probability sample designs: Using LAPOP's remote interview geo-locating of households in real-time (RIGHT) system (LAPOP Lab Methodological Note IMN004). LAPOP Lab.
- Pew Research Center. (2020). *Assessing the risks to online polls from bogus respondents*. Retrieved from https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2020/02/PM_02.18.20_dataquality_FULL.REPORT.pdf
- Porras, J., & English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. In Proceedings of the American Statistical Association, Section on Survey Research Methods (pp. 4223–4228). American Statistical Association.

- Pring, C. (2017). *People and corruption: Latin America and the Caribbean*. Berlin, Germany: Transparency International.
- Robbins, M. (2018). New frontiers in detecting fabrication. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods* (pp. 777–809). Hoboken, NJ: Wiley.
- Rosmansyah, Y., Santoso, I., Bani Hardi, A., Putri, A., & Sutikno, S. (2019). Detection of interview falsification in Statistics Indonesia's mobile survey. *International Journal on Electrical Engineering and Informatics*, 11(3), 474–484. doi: [10.15676/ijeei.2019.11.3.2](https://doi.org/10.15676/ijeei.2019.11.3.2)
- Sarracino, F., & Mikucka, M. (2017). Bias and efficiency loss in regression estimates due to duplicated observations: A Monte Carlo simulation. *Survey Research Methods*, 11(1), 17–44. doi: [10.18148/srm/2017.v11i1.7149](https://doi.org/10.18148/srm/2017.v11i1.7149)
- Schäfer, C., Schräpler, J.-P., Müller, K.-R., & Wagner, G. G. (2004). *Automatic identification of faked and fraudulent interviews in surveys by two different methods* (DIW Discussion Paper 441). German Institute of Economics. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.42515.de/dp441.pdf
- Schräpler, J.-P., & Wagner, G. G. (2003). Identification, characteristics and impact of faked interviews in surveys: An analysis by means of genuine fakes in the raw data of SOEP (IZA Discussion Paper 969). Institute for the Study of Labor. Retrieved from <http://hdl.handle.net/10419/20205>
- Seligson, M. A. (2007). The rise of populism and the left in Latin America. *Journal of Democracy*, 18(3), 81–95. doi: [10.1353/jod.2007.0057](https://doi.org/10.1353/jod.2007.0057)
- Shamir, M., & Sullivan, J. (1983). The political context of tolerance: The United States and Israel. *American Political Science Review*, 77(4), 911–928. doi: [10.2307/1957566](https://doi.org/10.2307/1957566)
- Silver, L., McRoy, M., Devlin, K., & Moynihan, P. (2019, May 14). *Who's home and who isn't? The challenges of conducting face-to-face interviews in Jordan*. Decoded/ Pew Research Center. Retrieved from <https://medium.com/pew-research-center-decoded/whos-home-and-who-isn-t-the-challenges-of-conducting-face-to-face-interviews-in-jordan-95a2a7081ded>
- Singer, M. M. (2018). Delegating away democracy: How good representation and policy success can undermine democratic legitimacy. *Comparative Political Studies*, 51(13), 1754–1788. doi: [10.1177/0010414018784054](https://doi.org/10.1177/0010414018784054)
- Slomczynski, K. M., Powalko, P., & Krauze, T. (2017). Non-unique records in international survey projects: The need for extending data quality control. *Survey Research Methods*, 11(1), 1–16. doi: [10.18148/srm/2017.v11i1.6557](https://doi.org/10.18148/srm/2017.v11i1.6557)
- Turner, C. F., Gribble, J. N., Al-Tayyib, A. A., & Chromy, J. R. (2002). Falsification in epidemiologic surveys: Detection and remediation. *Survey Research Methods Technical Paper on Health and Behavior Measurement*, 53. Research Triangle Institute
- Waller, L. G. (2013). Interviewing the surveyors: Factors which contribute to questionnaire falsification (curbstoning) among Jamaican field surveyors. *International Journal of Social Research Methodology*, 16(2), 155–164. doi: [10.1080/13645579.2012.687560](https://doi.org/10.1080/13645579.2012.687560)
- 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. doi: [10.3233/SJI-161012](https://doi.org/10.3233/SJI-161012)
- Winker, P., Kruse, K. W., Menold, N., & Landrock, U. (2015). Interviewer effects in real and falsified interviews: Results from a large scale experiment. *Statistical Journal of the IAOS*, 31(3), 423–434. doi: [10.1177/0759106317725640](https://doi.org/10.1177/0759106317725640)