Don't Get Duped: Fraud through Duplication in Public Opinion Surveys [1]

Noble Kuriakose[1] and Michael Robbins[2]

[1]SurveyMonkey
[2]Princeton University
[2]University of Michigan

December 12, 2015

[1] We thank Jon Cohen, Mingnan Liu, James Lo, Walter Mebane, and Jeremy Uecker for sharing their valuable insights in various stages of our research. We're also grateful for the feedback we received at the New England Chapter of the American Association of Public Opinion Researchers workshop on Fabrication in Survey Research, the American Political Science Association annual meeting, and the Total Survey Error conference. The views expressed in this paper are our own and do not necessarily reflect the views of our organizations.

Abstract

Fraud in survey research can take many forms, but a common form is through duplication of valid interviews. Duplication of a valid interview has a number of advantages: expected relationships between the variables will hold across the data set and, if done across a number of interviews, this approach can evade many standard techniques to detect fraud such as straight-lining analysis and the application of Benford's law. In this paper, we consider the likelihood of encountering near duplicates in survey data, suggest methods to fingerprint suspicious observations, report on our analysis of over 1,000 publicly available survey datasets and argue that nearly one in five widely used country-year surveys surveys from major international data sets have exact or near duplicates in excess of 5% of observations.

Introduction

Valid inferences in the social sciences depend on the quality of the data on which they are based. Unfortunately, we report that fraudulent data are potentially present in nearly 1 in 5 surveys from widely used cross-national datasets. In this analysis, we provide a short introduction to the problem of interviewer falsification, statistical evidence of falsified data, discuss alternative explanations for anomalous data, and lay out the implications of our findings.

Background on Fraud in Survey Data

As with any enterprise, data collectors hired for a project have an incentive to cheat to save time and money. Falsification can take many forms (AAPOR, 2003), but one of the most common and longest-running problems is cheating by interviewers (Bennett, 1948).

In response to falsification by firms or interviewers, researchers have developed techniques to detect these and other problems of cheating common to survey research by checking for straight-lining or applying Benford's law (Lawrence and Love, 2010; Porras and English, 2004; Schäfer et al., 2004; Biemer and Stokes, 1989). Yet, as researchers' techniques to detect falsification grow more sophisticated, so do the methods falsifiers will employ to conceal cheating (Scheuren, 2015).

A specific but until recently under-examined form of cheating is falsification through duplication of responses from valid interviews. Advances in statistical software have allowed researchers to checks for exact duplicates—meaning observations whose response pattern matches 100 percent with another observation—and remove questionable data. Yet, many datasets have exact duplicate observations and this standard vastly understates the problem by providing fraudsters with a simple solution: change the response to a single variable between a valid interview and a duplicated observation and the falsification remains undetected.

For dishonest firms and interviewers, duplication may be more attractive than simply inventing fake responses. If a dishonest firm carries out a sufficient number of interviews among a diverse—and reasonably representative—segment of the target population, then the results will generally yield both the expected distributions on known variables and the proper correlations between variables.[1] If the observations in this partial survey are duplicated one or more times, then the required survey sample size is reached at a substantially lower cost.

It can be accidental, but when done on a large scale, it is more likely to be the result of an intentional effort to save time and money in the data collection process. It is possible for either a single interviewer, supervisor, or someone in the head office of the firm to duplicate observations.[2]

Expected Levels of Near Duplicates

We define near duplicates as observations that share a high number of responses with another observation. We calculate the affinity of observations by determining the percent match, meaning the maximum percentage of variables that an observation shares with any other observation in the data set. In other words, in a data set with 100 substantive variables, if observation A shares 99 common responses with observation B but

[1] Although it is possible for an interviewer or firm to copy fabricated interviews, there are limited expected benefits to this approach. A key benefit of duplicating valid interviews is that relationships between variables remain consistent, patterns of variance are as expected, and the results are more likely to pass basic logic checks. Duplication of fabricated interviews is less likely to yield these outcomes.

[2] For the scope of our present analysis, the precise source of falsification is not central as regardless of the perpetrator the end result is the same. However, for those leading a survey project, detecting where falsification occurs is a major concern. With sufficient paradata, forensic analysis, and discussions with the local firm, it may be possible to identify precise source of the falsification which can be used to determine what actions should be taken to address the issue in a particular data set.

is not an exact duplicate for any other observation, it would be considered a 99 percent match. Meanwhile, if observation C shares 95 variables with observation A and fewer with all others, then it would be considered a 95 percent match. Thus, the process involves comparing each observation with every other observation in the data set and assigning a value representing the maximum percentage of variables on which it shares responses with any other observation. This process is performed by a publicly available Stata programed developed by the authors called *percentmatch*.

In order to determine the maximum expected percentage of responses that two observations are likely to share in un-doctored data, (a) we consider the literature on the reliability of beliefs among mass publics and (b) develop simulated survey data for analysis.

Reliability of Public Opinion

It has long been known that the majority of citizens do not have deeply held or consistent opinions about a wide range of political and social topics. Converse (1964) finds that even on highly controversial and well-publicized issues, the majority of respondents may answer as if by flipping a coin. Citizens are particularly inconsistent in terms of degree of belief, meaning that the difference between a response of 'agree' and 'strongly agree' to a question is somewhat random. Although the degree of consistency is a function of political sophistication, including a respondent's level of education, interest in politics and level of information, Converse shows there is great variation in the answers provided to the same questions by the same respondent over time.

Similarly, Zaller (1992) argues that few citizens are inherently ideological and all face competing concerns and interests when providing their opinions to survey questions. He theorizes that a respondent's provided answer is a function of these competing considerations and is moderated by their top-of-the-head response, meaning whatever happens to be on their mind at that very moment. Thus, interviewing the same person at different points in time will yield different response patterns.

One major implication of these findings is that it is extremely improbable that the same individual will provide the exact same responses to every question on a survey if she takes the same survey multiple times. This is especially true if the survey instrument is lengthy, covers a variety of topics, and responses include five- or seven-point scales. Since a respondent's answer to a given question represents a somewhat random draw from her distribution of opinions, in a series of repeated random draws for each question there is an extremely low probability of duplicating the exact response pattern.[3]

By extension, even if two individuals are highly similar in backgrounds and view, the likelihood of them providing exactly the same responses to a lengthy survey is infinitesimally small. Thus, the expectation of a response pattern approximating perfect duplicates between any two observations within a single data set is statistically indistinguishable from zero.

Simulated Data

To estimate the maximum expected match we should expect to see in survey data without falsification, we constructed 100,000 simulated datasets with the following characteristics: 1,000 observations, 100 variables, and random assignment of two distinct values for each cell (0/1).

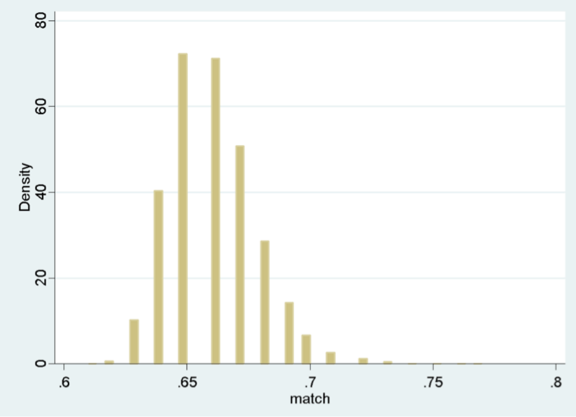
We chose to include only two distinct values in cells, meaning that the simulations have 200 potential responses across the 100 variables. By comparison, most major aca-

[3] Even if a respondent had a 95 percent likelihood of providing the same response to each item, which far exceeds the likelihoods suggested by Converse and Zaller, the probability that the exact same responses will be reproduced on a survey of 100 questions is less than 1 percent.

demic or policy survey projects use 4- or 5-point scales across many variables and also include ‘don’t know’ and ‘refuse’ as options, meaning there would generally be 500 or more potential response options across 100 variables. Thus, the results from the simulation offers a somewhat conservative test of the maximum percent match than might be expected between two observations in the survey.

After calculating the maximum percent match between observations and plotting the distributions, we found that the distribution closely resembled a Gumbel curve with an average mean of 0.66 ($\beta = 0.15$). Additionally, across the 100,000 simulated data sets, in no case did two observations have a maximum percent match that exceeded 85 percent.

Figure 1: Probability Density Function for Percent Match on Simulated Data



Since each response in the simulated data is completely independent and there is no assumed correlation between respondents, we conducted addition simulations on correlated data as a robustness check on the low likelihood of high percentage matches between observations. Simulated data with correlations were constructed with the following correlation structure, represented as a lower triangular matrix in table 1.

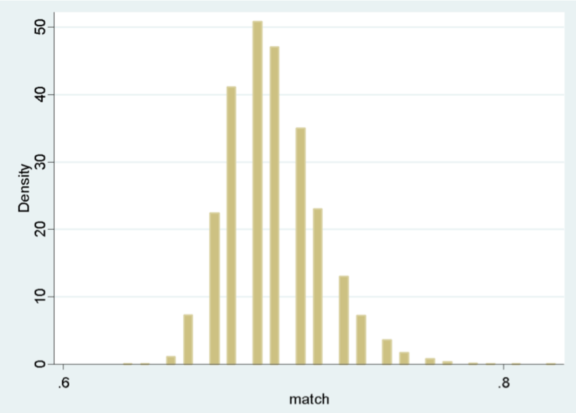
C at any intersection in the matrix is a uniformly distributed random number generated on the interval [-.9, .9]. The generated correlation matrix is then modified to be

Table 1: Simulated Correlation Matrix

	k_1	k_2	k_3	k_n
k_1	1			
k_2	$C_{k_2k_1}$	1		
k_3	$C_{k_3k_1}$	$C_{k_3k_2}$	1	
k_n	$C_{k_nk_1}$	$C_{k_nk_2}$	$C_{k_nk_3}$	1

positive semidefinite by setting negative eigenvalues to 0. Data that fit the correlation matrix specified were created for 1,000 observations and 100 variables (k). Values for each cell in the data were then recoded from real numbers [-6, 6] to two discrete values 0 or 1 based on the whether the real number was positive or negative. Although the recoding tempers the correlations between variables in the data, it more closely simulates real-world survey conditions, i.e. integer values. After 100 simulations following the steps above, we found that the highest values for percentage match were also less than 85 percent. Once again, when plotted the overall distribution approximates a Gumbel distribution.

Figure 2: Probability Density Function for Percent Match on Correlated Simulated Data



The results of simulated data suggest that the maximum percentage match between observations should not exceed 85 percent in un-doctored data.

As an check of these simulations, we compared the predictions with a number of leading academic surveys known for rigorous sample designs and oversight: the General Social Survey (1972-2014) and the American National Elections Study (1948-2012). None of these surveys contain any exact duplicates.[4] After accounting for skip patterns and early surveys that had fewer than 100 questions asked of the entire sample, only 35 observations out of more than 95,000 analyzed had a 85 percent or greater percent match (<0.05%).[5]

The Scope of the Problem

To assess the degree to which near duplicates are present in major international surveys, we analyzed a number of widely used and publicly available cross-national data sets. The surveys we analyze cover a number of major projects that appear widely in peer-reviewed journals as well as many policy papers and media outlets. All data sets that were analyzed were downloaded in 2015. In total, we analyzed 1,008 national surveys with more than 1.2 million observations, collected over a period of 35 years covering 154 countries, territories, or subregions.

Each survey in these projects is nationally representative or nearly nationally representative.[6] Most surveys have 1,000 or more respondents, although a select number have fewer in select countries. Each instrument is lengthy, covering 75 or more questions or more on a range of topics.[7] Multiple modes were employed to collect the data including CATI, CAPI, and PAPI.[8] Among those collected using clustered samples, the cluster size

[4] The one exception being the 1980 American National Election Study with 30 cases that were purposely duplicated as explained in the codebook.

[5] Given the short survey instrument administered by the ANES prior to 1964, our analysis is limited to surveys from 1964-2012.

[6] A small number of surveys are representative of a region or sub-region of a country.

[7] Based on the criteria discussed below, not all substantive questions are included in the analysis. As such, the final number of variables in specific country-year surveys in the analysis that follows is often less than 75. We implemented a cutoff for analysis at more than 35 eligible questions. A discussion of the potential effects of the number of question number on the expected number of observations with a high percent match is detailed later in the paper.

[8] Based on the methodological statements provided by the projects, it is not always possible to determine

is generally 12 or fewer.

For each survey, we examined the distribution of maximum percent match for substantive variables for every unique country-year to determine if there is a significant likelihood of substantial data falsification via duplication.[9] We use the term "potential falsification" intentionally. It is possible that in some instances what appears to be falsification is in fact the result of some other artifact in the data set, such as the intentional duplication of 30 observations in the 1980 ANES, or another potential explanation.[10]

To further guard against false positives, meaning near duplicates that are the result of some factor other than falsification, we take additional steps in the analysis that follows. First, we removed any variable where 10 percent or more observations had a missing value. This approach eliminates variables that may be part of a skip pattern, such as in a split sample battery. Second, we removed any observation where 25 percent of more of variables were missing. This approach minimizes the risk of having break offs skew the analysis and overstate the level of similarity between observations.

Drawing on results of simulations, we examine the degree to which near duplicates are present in unique country years from the surveys listed above. We consider two benchmarks: those that fall above 10 percent and 5 percent. The former barrier has a significant likelihood of biasing statistical analysis, including by artificially increasing statistical power and decreasing the variance, resulting in smaller estimated confidence intervals for point estimates. Similarly, the latter threshold has a moderate likelihood of biasing analyses that

the exact mode for each country-year. However, the data set includes surveys done by each of these three modes. Although CATI or CAPI may reduce the likelihood of falsification by interviewers, it does not mean it would be impossible for a firm or interviewer to falsify data in this manner. Understanding the differences in incident rates between modes would be a welcome course for future research.

[9] Our analysis has shown that exact duplicates are commonly found among substantive variables but less commonly found in geographic or demographic variables. Unlike substantive variables, geographic variables must match the pre-determined sampling plan while demographic variables should approximate population parameters. Thus, the benefits to a dishonest firm or interviewer of duplicating these items are lower than for substantive variables. See also Waller (2013) and Robbins (2015).

[10] The ANES is not one of the data sets included in the analysis of the 1,008 country-year data sets.

include the likely falsified observations. While analyses with lower levels of near duplicates are also likely falsified and should be accounted for by researchers, they have a relatively low likelihood of significantly biasing conclusions.

Table 2: Level of Duplication Across 1,008 Country-Year-Surveys

Degree of Likely Falsification	**Percentage of Surveys**
No cases	35.8%
<5%	46.8%
5% to <10%	7.2%
≥ 10%	10.1%

We find that nearly one in five country-year surveys in publicly available datasets included in our analysis has a level of near (or full) duplication of 5 percent or greater. These results imply that duplicates and near-duplicates present a prevalent problem that has been largely undetected to date.[11].

Additionally, there are certain factors that make a higher level of near duplicates more likely in a survey. For example, additional analysis reveals that near duplicates are much more likely to be found in surveys conducted in non-OECD countries compared with members of this club of wealthy countries (see table 3).[12] In non-OECD countries the rate at which exact or near duplicates exceed 10 percent of all observations is 15.3 percent compared with just 2.0 percent in OECD countries. Additionally, the percentage of surveys with no observations that match another at the 85 percent level or higher in OECD countries is 52.6 percent compared with just 17.3 percent in non-OECD countries. Overall, the chance of having a level of near duplicates (≥85 percent match) that exceeds 5 percent in non-OECD countries is 21.5 percentage points higher than in OECD countries.

[11] A notable exception is Koczela et al. 2015 and a number of recent workshops sponsored by NEAAPOR and the Washington Statistical Society

[12] Out of the 1,008 country-year data sets, 350 were conducted in current OECD countries and 658 are from non-OECD countries.

Table 3: Level of Duplication by OECD Status

Degree of Likely Falsification	**OECD**	**Non-OECD**	***Difference***
No cases	52.6%	17.3%	*+35.2 pts.*
<5%	42.9%	56.5%	*-13.7 pts.*
5% to <10%	2.6%	10.8%	*-8.2 pts.*
$\geq$ 10%	2.0%	15.3%	*-13.3 pts.*

Alternative Explanations

General Hypotheses

A number of potential objections could be raised that might explain these results. Some possibilities include straight-lining by the respondent as a form of satisficing or high levels of non-response. If these possibilities were sufficient explanations, then nearly all surveys in the sample should exhibit a significant number of near duplicates for lengthy survey instruments. The fact that two-thirds of surveys, including surveys from each series, do not exhibit high levels of near duplicates calls into question both of these possibilities. Moreover, even if one of these behaviors explained the high percent match between two observations, this behavior does not yield a quality interview and an analyst may want to consider discarding it from their analysis.

Identical Neighbor Hypothesis

A second alternative hypothesis is that our simulations have underestimated the correlation between either variables in survey or between individuals in the population more generally. We refer to this as the identical neighbor explanation.

To evaluate this possibility, we analyzed a number of surveys with a limited target population over a concentrated geographical area on a relatively limited range of topics.

If there were indeed identical neighbors, we would expect to find more near duplicates in such surveys than in nationally representative ones like those we analyzed above.

A number of publicly available surveys of small areas exist, but a long-standing survey of a single metropolitan area in the US is of particular use for this purpose. This series of surveys, carried out between 1951 and 2004, covers three counties. The central topic of the survey varies by year, but most focus narrowly on a single issue.

In the early 2000s this survey project focused on the experience of a relatively small ethnic minority population living in the area.[13] In this survey, 1.1 percent of observations (11 of 1,016) have a percent match that exceeds 85 percent and none exceeds 87.4 percent.[14] Thus, even in a survey of a unique sub-population in a geographically concentrated area, the maximum percent match closely approximates expectations from the computer simulations. Results for other years of this survey project are similar.

Similar results are found for surveys as diverse as a survey from the mid-1980s on the presidential election in a small midwestern town, a survey on views toward the performing arts in moderate-sized US town, a survey on political economy in four counties in a country in East Asia, and a survey of attitudes toward the police in a New York county. All are examples of small geography surveys with almost no near-duplicates.[15]

A second test to evaluate the plausibility of the identical neighbor explanation is to compare similar surveys done in the same country. If near duplicate observations are present because respondents in the country are naturally more alike, all similar surveys in the country should find these patterns. We analyzed data from two surveys from a North African country carried out less than a year apart, which used nearly identical sampling

[13] Based on the 2000 census figures, the population of this ethnic minority for the entire state was slightly more than 100,000, making the target population for the study less than this total (de la Cruz and Brittingham, 2003).

[14] For additional details on this survey see appendix A.

[15] The survey from East Asia has a small percentage observations (3%) with a percent match exceeding 85 percent. In the performing arts survey, 0.2 percent share a percent match that slightly exceeds 85 percent.

methodologies and carried out by similar research teams but by different organizations.

We find that the two surveys have drastically patterns of similarity between observations. In the first survey 2 percent of observations have a percent match that exceeds 85 percent compared with the second where nearly 1 in 5 observations (18%) exceeds an 85 percent match.[16]

In sum, the hypothesis that the high concentrations of near duplicates are the result of higher degrees of correlation between the variables or between respondents has little basis for support. Many surveys of highly concentrated populations on variables that have a relatively high expected correlation rarely violate a maximum percent match of 85 percent. Moreover, these results are not linked with certain countries, but rather are more closely linked with specific surveys.

Number of Questions Hypothesis

A third alternative hypothesis is that the baseline threshold developed in this paper is inappropriate given variations in the number of questions in a survey and the number of response options. After all, the simulations described above have assumed 100 questions. Running similar simulations with only 10 questions and with two response options each would yield a distribution that had many observations in excess of an 85 percent match and likely many exact duplicate observations.

The statistical probability of finding exact or near duplicates is to some degree related to the number of questions, in addition to the true (but unknown) distribution of responses across question. Asking Americans if they support freedom would yield a significantly different distribution than asking them about their views of trading basic freedoms in the name of achieving greater security. Moreover, the true distribution is based on the number

[16] A more detailed analysis of these two surveys can be found in appendix A.

of response options yielding an additional complication.[17] The simulations and analysis on high quality data from real surveys yielded a general threshold of 85 percent, but perhaps this assumption leads to a significant number of false positives where there is no reason to suspect likely falsification via duplication in the analysis of the more than 1,000 surveys.

Computer simulations, especially given the unknown true distribution of responses, are likely to have limited value in establishing the expected number of near duplicates. Empirical analysis is better suited to determine the degree to which the percentage of observations with a high percent match is related to the number of eligible questions in the survey. Again, our initial computer simulations were based on 100 questions and 1,000 respondents in the goals of approximating the sample size and number of questions in an 'average' international survey project.

To evaluate if there is a significant relationship, we divided the full sample of over 1,000 surveys into two subsamples. The first included 275 surveys with the total number of questions ranging from 36 to 74 based on our stated criteria for eligible questions for analysis.[18] The median number of questions in this subset of surveys is 60. The second sample includes 733 surveys where the total number of eligible questions ranges from 75 to 347. The median number of questions in the second subset is 160, meaning there are more than two-and-a-half times the number of questions in the second sample compared with the first.

The expectation based on this hypothesis is that because of a substantially smaller number of questions in the first subset of country-year surveys, the percentage of percent

[17] The relationship between number of response options and the true distribution for a question are also challenging to estimate given diminishing returns. A five-point scale will yield greater variation than two response options, but a 100-point scale has a limited number of commonly chosen values, often yielding little more variation than a 10-point scale. Thus, modeling the relationship between these two variables would require significant assumptions within the model.

[18] These criteria eliminate questions with missing data for more than 10 percent of respondents or where the level of 'don't know' and 'refused' exceeds 20 percent.

matches that exceed 85 percent should be significantly higher than in the second subset of country year surveys. In fact, no such relationship is observed. The mean percent match in the surveys in the first subset (36-74 questions) is 2.8 percent compared with 3.9 percent in the second subset (75+). It also appears unlikely that this result is a function of the sample size given that the median sample size in the first subset is 1,010 compared with 1,200 in the second subset. Thus, despite expectations, empirical evidence reveals that it is unlikely that the overall findings are significantly biased due to variation in the number of questions across surveys.

Table 4: Percent Match by Number of Questions

Sample Subset	**Number of Surveys**	**Median Number of Questions**	**Mean Percent Match**	**Median N**
36-74 Questions	275	60	2.8%	1,010
75+ Questions	733	160	3.9%	1,200

Additional Test

A second key finding from the simulations is that the overall distribution of the maximum percent match should resemble a Gumbel distribution. Our analysis of a number of leading surveys projects, particularly those from the US and other OECD countries, closely resemble a Gumbel distribution. Additionally, even geographically concentrated surveys such as the survey of a small ethnic minority in a single metropolitan area closely approximate a Gumbel distribution (see appendix A). Plotting the maximum percent match yields an additional tool for researchers interested in improving data quality. Deviations from a Gumbel—when skip patterns, split ballot questionnaires, and break offs are controlled for—suggest there may be problems with falsification with duplication even when relatively few observations exceed a maximum percent match of 85 percent. One such instance comes from a survey from a non-OECD country.

In this case, the initial test suggests a moderate likelihood of data falsification via duplication with 5.6 percent of all observations having a percent match that exceeds 85 percent. Although the general distribution approximates a Gumbel, on the right tail there are small but discernible peaks when percent match exceeds roughly 80 percent (see figure 3). As a result, we closely examined all observations with a percent match exceeding 80 percent.

Figure 3: Deviation from a Gumbel Distribution

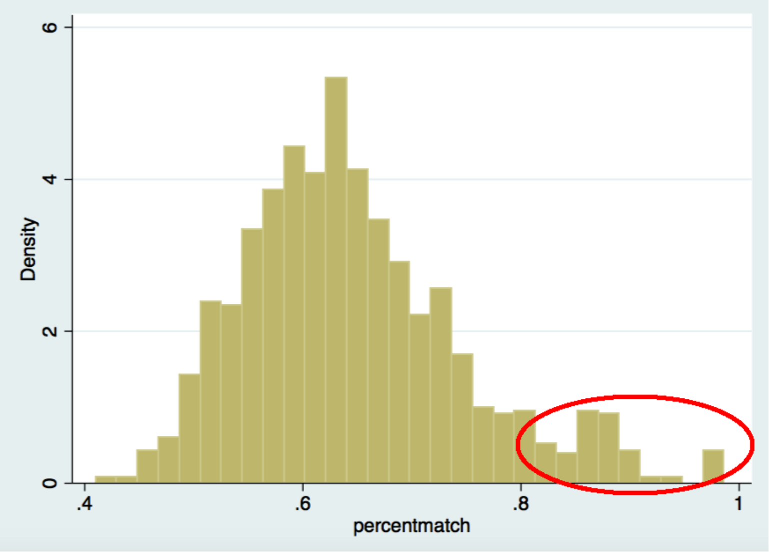


Table 4 presents instances when percent match exceeds 80 percent by interviewer.[19] For eight of the ten interviewers, only a small minority of their interviews fall into this range. However, for interviewers 8 and 9, nearly half exceed 80 percent.

More detailed analysis of the response patterns of these interviewers yielded a clear instance of suspected fraud. In a country where party identification tends to be weak, interviewers 8 and 9 had a very high level of affiliation for the main Islamist party in the

[19] The table includes only interviewers with observations with a percent match ≥80%.

Table 5: Number of Respondents by Interviewer and Percent Match

Interviewer	*% Match < 80%*	*% Match ≥ 80%*
1	41	9
2	31	3
3	28	2
4	39	11
5	45	5
6	91	4
7	52	31
8	21	19
9	23	17
10	44	6

country (see table 6). Similarly, support for political Islam was extremely high among their respondents. Notably, the region covered by these interviewers is not an area where the main Islamist party has performed particularly well in elections. However, it was possible to identify a second survey that employed these interviewers which included topics on political Islam. Again, among respondents of these two interviewers, support for political Islam exceeded the national average, strongly suggesting that the two were falsifying responses, at least on these variables. Thus, even in cases where the maximum percent match does not exceed 85 percent, it may be possible to identify suspect observations based on deviations from the expected distribution.

Table 6: Support for Islamist Party by Interviewer

Interviewer	*Average % Islamist Support*	*Max % Islamist Support*
8 & 9	66.3	77.5
All others	10.1	24.1

Limitations

There are two important limitations for this analysis. First, examining the maximum percent match is likely to perform better a with a longer survey instrument, and the distribution is more likely to approximate a Gumbel with a larger-n survey . Small surveys with only a few questions are less suitable for analysis with this tool.

Second, this approach to identifying fraudulent observations in surveys does not necessarily translate well outside the world of social science surveys and particularly public opinion surveys. For customer satisfaction or employee engagement surveys, where every question is often, in effect, an item on a single scale, the share of respondents marking the highest or lowest available values for every question will be statistically anomalous.[20] Workers who are pleased (or displeased) with their work environment are likely to provide similarly high (or low) responses across many questions. Part of the reason near duplicates are less likely in social science public opinion polling is the diversity of questions and topics.

Conclusion

Falsification via duplication is an especially pernicious problem because it can produce survey data that appears to be valid. When undetected, it can significantly bias the analyses that make use of these data. Artificially increasing the number of observations has significant implications for statistical analysis; duplicate variables artificially increase statistical power and decrease the variance, resulting in smaller estimated confidence intervals for point estimates. Removing the duplicated observations means fewer significant correlations due to larger confidence intervals.

Although we encourage survey researchers to look at their past surveys to ensure data

[20] An examination of a large number of unpublished surveys that fit this description support this assertion.

meet the highest quality standards, our publicly available *percentmatch* program provides its greatest value for future work. If rigorously applied by researchers, falsification will become more difficult and costly for firms and interviewers. Raising the costs makes it more likely that data quality will be higher and provide more accurate representations of public opinion.

This was our goal in creating the program and making it widely available free of charge. Survey vendors and primary investigators should use this tool and our method to flag potentially problematic observations. Further analysis should be carried out to definitively determine the sources of near duplicates, whether they be innocuous or the result of fraud.

Finally, the issues we discuss in the paper present compelling reasons for academic journal editors to require article submitters to make survey microdata available to reviewers or an independent entity. Publications of findings on datasets that have not been scrutinized for quality by programs like ours may not hold up after accounting for fraud or other errors. Ultimately, oversight and transparency are key to ensuring data quality.

Note: Our Stata program to identify near duplicates, *percentmatch*, can be downloaded from SSC (the Statistical Software Components archive). Cite use of the program or our method by referencing this paper.

References

AAPOR. 2003. "Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effect." *AAPOR.org* .

Bennett, Archibald S. 1948. "Toward a solution of the "cheater problem" among part-time research investigators." *Journal of Marketing* 2:470–474.

Biemer, Paul P. and S. Lynne Stokes. 1989. "The Optimal Design of Quality Control Sample to Detect Interviewer Cheating." *Journal of Official Statistics* 5(1):23–29.

Converse, Philip. 1964. *The Nature of Belief Systems in Mass Publics.* Vol. Ideology and Discontent The Free Press of Glencoe.

de la Cruz, G. Patricia and Angela Brittingham. 2003. "The Arab Population: 2000." *Census 2000 Brief* .

Koczela, Steve, Cathy Furlong, Jaki McCarthy and Ali Mushtaq. 2015. "Curbstoning and beyond: Confronting data fabrication in survey research." *Statistical Journal of the IAOS* 31:413–422.

Lawrence, Chriselle and Elizabeth Love. 2010. "Characteristics of Falsified Interviews." *JSM Proceedings, Survey Research Methods Section* pp. 4771–4780.

Porras, Javier and Ned English. 2004. "Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys." *JSM Proceedings, Survey Research Methods Section* pp. 4223–4228.

Robbins, Michael. 2015. Preventing Data Falsificaiton in Survey Research: Lessons from the Arab Barometer. Presented at New Frontiers in Preventing, Detecting, and

Remediating Fabrication in Survey Research. Conferences hosted by NEAAPOR and Harvard Program on Survey Research, Cambridge, MA, February 13.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller and Gert G. Wagner. 2004. “Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methods.” *JSM Proceedings, Survey Research Methods Section* pp. 4318–4325.

Scheuren, Fritz. 2015. Comments at New Frontiers in Preventing, Detecting, and Remediating Fabrication in Survey Research. Conferences hosted by NEAAPOR and Harvard Program on Survey Research, Cambridge, MA, February 13.

Waller, Lloyd George. 2013. “Interviewing the Surveyors: Factors which contribute to questionnaire falsification (curbstoning) among Jamaican field surveyors.” *International Journal of Social Research Methodology* 19(2):155–164.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion.* Cambridge University Press.

Appendix A: Select Percent Match Distributions and Discussion

Figure 4: OECD Country Survey 1

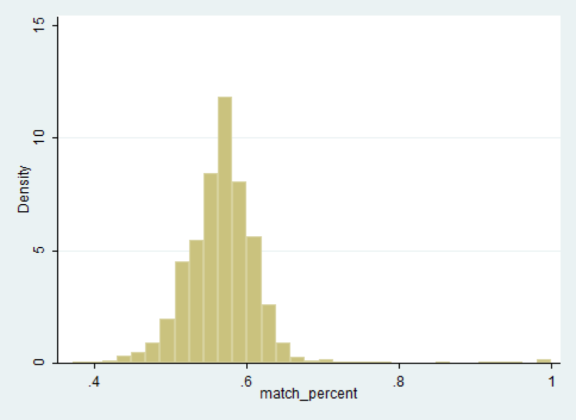
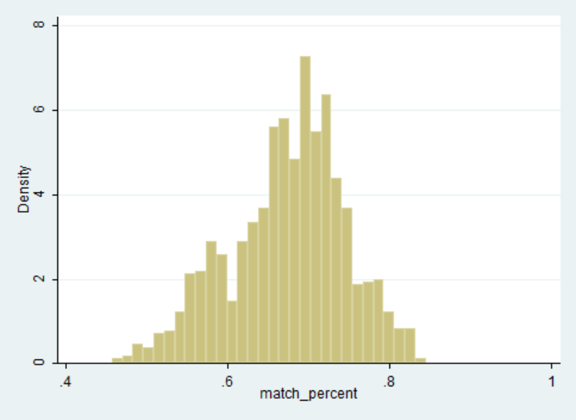


Figure 5: Non-OECD Country Survey 1



These two recent surveys closely approximate a Gumbel distribution with a modal maximum percent match of roughly 60 percent. There are a small number of exact duplicates in the the first survey while there are two observations with a maximum percent match that slightly exceeds 85 percent in the second survey.

Figure 6: Non-OECD Country Survey 2

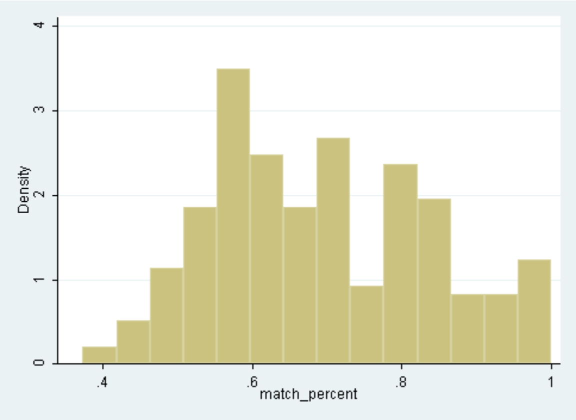
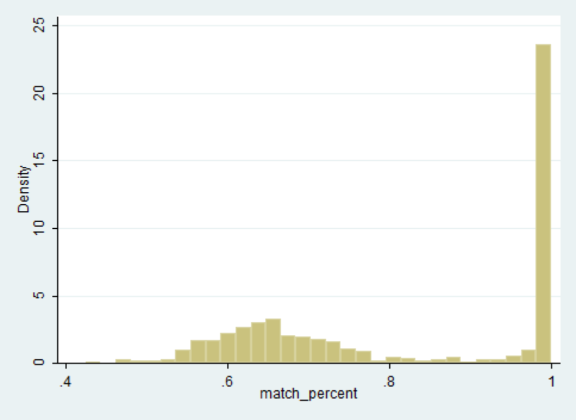
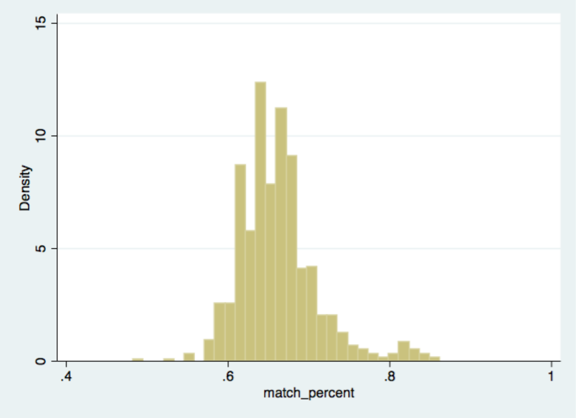


Figure 7: Non-OECD Country Survey 3



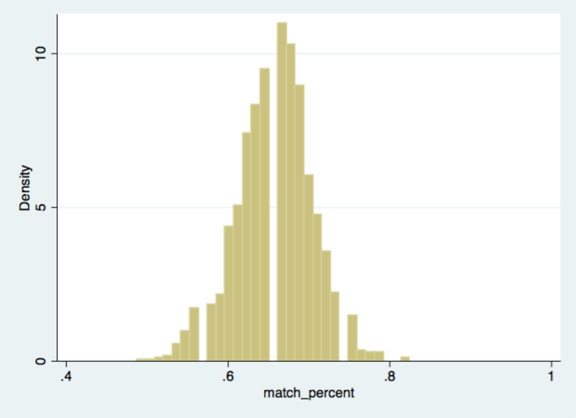
Nearly half of the observations from these two surveys from non-OECD countries have a maximum percent match that exceeds 85 percent. Both are also departures from the expected Gumbel distribution. In the first survey, 26 percent of observations are a 100 percent match on substantive variables and 54 percent exceed 85 percent suggesting likely falsification by the local firm. In the second survey, 39 percent of observations are a 100 percent match on substantive variables and 49 percent of observations have a percent match that exceeds 85 percent.

Figure 8: Ethnic Minority Survey in Three US Counties Percent Match Distribution



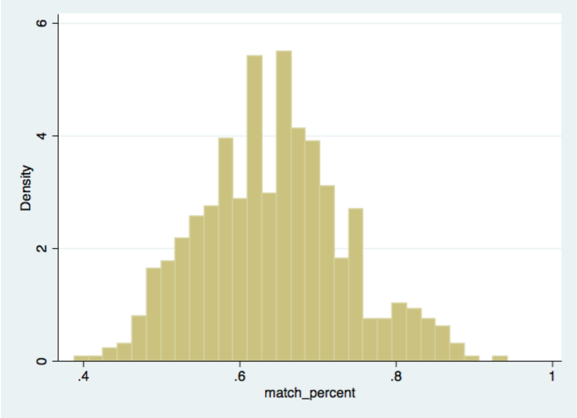
The distribution of the maximum percent match for this survey of a small ethnic minority from a single US metropolitan area generally approximates a Gumbel distribution with a mode of 70 percent, which is only 4 percent higher than the mode from the computer simulations. Using our criteria for eligible questions, only 2 of 1,016 observations have a percent match of 85 percent or higher. Thus, even in a survey of a unique sub-population in a geographically concentrated area there are few observations with a percent match exceeding 85 percent.

Figure 9: Presidential Survey in Midwestern Town Percent Match Distribution



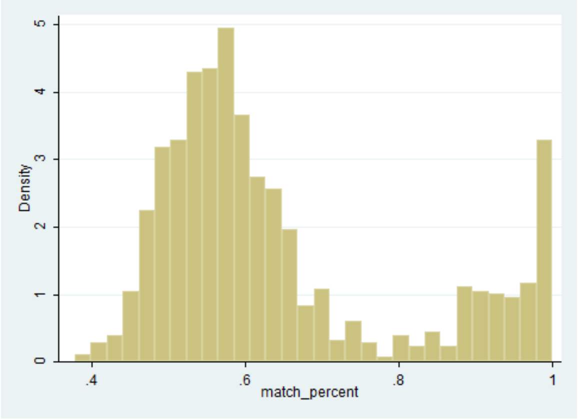
The distribution of the maximum percent match for a survey carried out in a small midwestern town that had a population of just over 100,000 residents in the mid-1980s. The first wave of the survey (n=1,488) was distributed across 16 clusters (neighborhoods) and the survey instrument focused on political attitudes. The distribution closely approximates a Gumbel with a mode of roughly 65 percent. Despite a highly clustered design in a relatively small town and only 80 questions centered on a U.S. presidential election, the maximum percent match for all observations never exceeds 85 percent.

Figure 10: African Country Percent Match Distribution Survey 1



This survey was conducted recently in Africa and very roughly approximates a Gumbel distribution with a mode of approximately 65 percent. Additionally, there are 2 percent of observations with a percent match that exceeds 85 percent. These observations are higher than what would be expected, suggesting there may be some minor issues among a small number of observations in the data set.

Figure 11: African Country Percent Match Distribution Survey 2



A second survey was conducted less than a year later using a nearly identical sampling methodology to the first survey, including its sampling plan. The primary difference was the team that led the survey. By comparison, the overall distribution is far from a Gumbel distribution. The distribution is non-monotonic to the right of the mode, (approximately 65 percent). Instead, there is a second peak on the right tail. Of the nearly 1,200 observations that were included in the analysis, (18%) have a percent match higher than 85 percent. Notably, only 32 total are exact duplicates (100% match) but 109 additional observations have a percent match that exceeds 95 percent.