Falsification in Surveys: Detecting Near Duplicate Observations[1]

Noble Kuriakose
Research Scientist
SurveyMonkey
noblek@surveymonkey.com

Michael Robbins
Research Director
Arab Barometer
robbinmd@umich.edu

March 18, 2015

[1]The authors' names are listed alphabetically.

Abstract

One of the primary concerns of survey researchers is that faraway interviewers faithfully poll ordinary citizens and accurately report their responses so they can be analyzed with confidence. The use of duplicate observations has long been employed by unscrupulous vendors and field workers as a way to artificially boost the number of cases delivered to a researcher, especially in international survey work. If duplicates are based closely on data collected from valid interviews, the benefits of falsification are clear: the observations correlate as expected and raise no red flags. Most data programs now include standard tests that can easily locate exact duplicates, which can be removed from data sets before analysis. However what is more difficult is detecting near duplicates, falsified observations where only a limited number of variables have been copied from genuine responses in order to slip past tests for exact duplicates.

Our paper first establishes the degree to which near duplicates are present in a number of major survey research projects, and shows this problem is widespread. We demonstrate why this form of falsification often goes unnoticed, and offer some guidelines for determining the likelihood a data set includes near duplicates. Finally, we detail a new Stata program that tests for near duplicates and can be used as a diagnostic to evaluate a data set's quality.

Introduction

Valid inferences in the social sciences depend on the quality of the data on which they are based. For survey data, this means ensuring that interviews were conducted according to the research methods set out by the survey. Yet, as with any enterprise, data collectors have an incentive to cheat. Thus, data falsification remains a primary concern in the collection of public opinion surveys.

Falsification can take many forms (AAPOR, 2003), but one of the most common and longest-running problems is cheating by interviewers (Bennett, 1948). A specific but underexamined form of cheating by interviewers is the falsification through duplication of responses from valid interviews. This form is an especially pernicious problem because it can produce survey data that appears to be valid and when undetected can significantly bias the analyses that make use of these data.

In the past, the most common checks for falsification focused on locating exact duplicates, in which one observation is identical on all variables to a second observation. As checks for exact duplicates have become routine using basic statistical software, duplication has seemed easier to discover. However, a brief review of a number of publicly available data sets reveals that falsification via duplication remains extensive, although it has gone mostly undetected to date because it has emerged in a new, tougher-to-identify form: falsification through near duplicates. Near duplicates are duplicate observations in which responses for only a small number of variables are changed in an effort to prevent them being detected by standard tests for exact duplicates. In particular, we show that duplication remains relatively pervasive across substantive or attitudinal variables within a survey.

Falsification via duplication can, and often does, lead to misleading findings. 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. Moreover, duplication is more likely to be implemented for populations that are harder to reach including those in remote areas or younger respondents. By systematically excluding these individuals from the survey, point estimates are also likely to be biased. Indeed, eliminating duplicate and near duplicate observations from analysis is imperative to ensuring valid inferences.

In this paper we begin by outlining the basic logic of duplicating observations, examining prior attempts to detect duplicate observations, explaining why it remains one of the most effective strategies for data falsification, and discussing when it is most likely to happen. In the next section we detail our newly developed Stata program designed to more effectively test for duplicate observations. In the third section we use the program to examine the degree to which duplicates are present in a number of publicly available datasets and show how the program can be used to identify different types of duplication. We conclude with recommendations to researchers for using the program.

The Logic of Duplication

To properly root out falsified data, it is important to understand when and why falsification occurs. 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.

One of the biggest challenges in data falsification is ensuring the proper correlations between different variables in a data set. If results are entered at random, they risk differing significantly from other surveys conducted among the same target population. Moreover,

the internal logic of the survey is also unlikely to hold; known correlations across key variables can disappear if the survey is filled out more-or-less at random. For example, one may expect to see that those with more conservative attitudes on morality also have more conservative beliefs about gender roles. Randomly falsified results might not contain that correlation.

This makes duplication 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 national population, then the results will generally yield both the expected distributions on most variables and the proper correlations between variables. 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.

Some variables are more likely to be falsified in this manner than others. Namely, substantive variables, meaning attitudinal questions, are more likely to be duplicated. These variables generally present new and unknown (or at least uncertain) distributions for the survey research firm or interviewer. Without conducting real interviews, there is an almost certain probability that one or more of the variables will not yield the expected distribution or relationships with other variables.

Although duplicates may affect any type of survey, they are more likely to occur in face-to-face surveys than in those done by other modes. Face-to-face surveys are both the most costly to carry out and also the most difficult to oversee. This is especially true for surveys being taken in foreign countries where the presence of supervising foreign researchers around the sampling point may bias results.

Especially before recent advancements with the use of computer-assisted personal interviewing (CAPI), it remained difficult to ensure interviewers were faithfully carrying out their interviews. Survey researchers were forced to rely on the trustworthiness of the

firm and the firm's technical capacity to ensure falsification was not taking place in real time. In most cases, only upon receipt of the final data could survey researchers begin to assess the degree to which falsification was likely in the results. Timers and other safeguards such as geotagging or recording snippets can be used with CAPI in face-to-face interviews. However, especially in international surveys, the use of CAPI is not always feasible, and even then it is not a perfect safeguard against falsification.

By contrast, incentives to falsify phone interviews are lower, due both to the cheaper cost and simpler oversight requirements. Interviews by phone can be recorded relatively easily, and interviewers can be monitored in person, without biasing the respondent's replies.

Measuring Attitudes

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. He found significant portions of citizens do not hold consistent opinions, especially in terms of degree of belief. 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 answer provided 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, the same individual is likely to provide different

responses to a survey depending on his or her unique recent experiences or information.

Considered together, Converse and Zaller imply that it is highly unlikely that even if the same person took a survey twice that he or she would provide the identical set of responses. This is particularly true if the survey was long and responses were scales with more than two possible options, given that attitudes are not firmly fixed. Thus, a respondent's answer to a given question represents a somewhat random draw from his or her distribution of opinions on the subject.

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. Given that in a standard survey relatively few individuals share the same age, gender, education, income level, and geographic area, the odds of receiving the same response pattern from multiple respondents within the same survey is even lower. In sum, the expectation of a very high percent match between any two observations within a single data set is statistically indistinguishable from zero.[1]

Linkage Analysis

There is a body of knowledge in computer science that addresses duplication: linkage analysis. Linkage analysis attempts to solve two issues: removing duplicate records from a table and matching records across tables. In both cases, the key is developing algorithms to identify the same records despite not being duplicates. For example, when examined individually, it is clear that the addresses 123 3rd St. and 123 Third Street indicate the

[1]There is likely to be significantly less variation in response patterns on a short survey that covers a single latent concept among respondents who are not independent. Such examples may include a workplace survey or student course evaluations. Workers who are pleased (or displeased) with their work environment are likely to provide similarly high (or low) responses across many questions. Moreover, since workers are likely to have received similar treatments and also discussed issues related to workplace quality prior to taking the survey, they are significantly more likely to provide highly similar response patterns.

same place. Or, that Jon Smith and Jonathan Smith could refer to the same person. Linkage analysis attempts to make that connection mechanically.

A review of the linkages literature identifies a number of strategies that seek to detect duplicate observations. Most of these strategies target string variables, attempting to determine the degree of similarity between text entries. Examples that seek to identify character-based similarities—meaning how similar two entries are to one another—include edit distance (Navarro, 2001), affine gap distance (Smith and Waterman, 1981), Smith-Waterman distance (Waterman, Smith and Beyer, 1976), Jaro distance metric (Jaro, 1976), and Q-Grams (Ullmann, 1977; Ukkonen, 1992). Other algorithms focus on the similarity of entries regardless of order by targeting strings of text that are similar between two observations. Examples of this strategy include atomic strings (Monge and Elkan, 1996) and WHIRL (Cohen, 1998). Other approaches for dealing with similarities of string variables include phonetic comparisons (Taft, 1970).

A common challenge in computer science and population research is to match multiple fields across records (Elmagarmid, Ipeirotis and Verykios, 2007). Through a process of Bayesian learning based on one (or more) of the aforementioned strategies, these algorithms develop methods that will estimate the probability that any two records are a match. Ultimately, this process produces a likelihood ratio that any two observations are duplicates.

The linkages approach assumes records are valid, especially when joining across tables or data sets. Within the world of survey research, the basic assumption about duplicates is the reverse; all but one of near duplicate observations are likely invalid. Thus, although we follow a similar process to the linkages literature, we depart from it in our attempt to identify fraud.

A second key departure we make from much of the linkages literature is our focus

on numeric values as opposed to string variables. Although string variables may also be identical, detecting these duplicates is far more challenging than detecting identical matches on fixed response items that are assigned a numeric value. We also do not focus on any specific variables, but examine the overall match across the full range of variables.

A third departure we make from the existing literature is we do not focus on long strings of duplicates. Others, such as Mushtaq (2014), have sought to evaluate the likelihood of falsification by detecting long strings of duplicate response patterns across consecutive variables. This approach depends on the order of variables to detect partial matches throughout the data set. We argue that order is less important than the overall percent match across all observations.[2] In our approach, even if every n^{th} variable is randomized, it will still produce a higher than expected percent match for the duplicated observation.[3]

Detecting Duplication

Tests for exact duplicates have become standard in recent years. Statistical software such as SPSS and Stata have simple packages to test for exact duplicates over the full or user-specified range of variables. Yet these basic safeguards significantly understate the degree to which duplication is likely to be present in a data set for two reasons.

First, not all variables are equally likely to be falsified via duplication (Waller, 2013). For example, the incentives to falsify demographic variables via duplication are low. These variables must approximate certain parameters such as census data or other reliable demographic estimates. Falsification via duplication is more likely to miss these targets than

[2]This approach yields a clear benefit for breakoffs or other instances where a part of the survey was skipped completely and the results copied over from another observation. However, even in this case, the assumption that consecutive variables match exactly is limiting, given that an unscrupulous interviewer or firm could modify every n^{th} variable in an attempt to conceal the falsification.

[3]Alternatively, if too many variables are falsified at random, then it will weaken the expected correlations across the substantive variables within the dataset.

fabricating demographic variables through other means would be.

Similarly, geographic variables provide few incentives for falsification via duplication. These variables must match the sampling frame, which cannot be achieved through duplication. Meanwhile, paradata does not need to correlate closely with any of the substantive variables. Instead, to pass standard quality tests, it must simply be relatively clean, meaning, for example, interview times cannot overlap, interviews should be randomized throughout the day, and interviewers cannot travel across the country in the same day. In sum, the process of duplication yields few benefits for falsifying these data.

Instead, as noted, the main gains from falsification are ensuring that attitudinal or behavioral variables will approximate expected point estimates and correlate as expected across items. Although complex algorithms may produce such results, many times it may be easier and relatively cost-effective for the unscrupulous firm to carry out a number of interviews and duplicate them to achieve this result. Thus, it is most important to test for duplicates among the substantive variables in a data set.

The second significant limitation of testing for duplicates, even among substantive variables, is that these tests rely on perfect matches. If a single variable is changed between two observations, standard programs do not recognize it at as an exact match. Although testing for a range of variables within the data is also possible, modifying the responses to a handful of variables for each observation across the survey may go undetected. If modified variables are randomized throughout the data set, these tests may not detect the full scope of the problem. A better testing method is a form of linkage analysis that examines the degree to which any two observations within the data set share common response patterns.

To detect partial duplicates, we propose determining the maximum percentage of variables that an observations 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 is not a perfect duplicate for any other observation, it would be considered a 99% match. Meanwhile, if observation C shares 95 variables with observation A and fewer with all others, then it would be considered a 95% match. Thus, each observation is compared with every other observation and assigned a value representing the maximum percentage of variables on which it shares responses with any other observation.

We argue that plotting the density function of the maximum percentage match for each observation should yield a Gumbel distribution. We know the ideal distribution and the range of likely maximum percentage values because we analyzed both simulated data and high quality survey data from the United States.

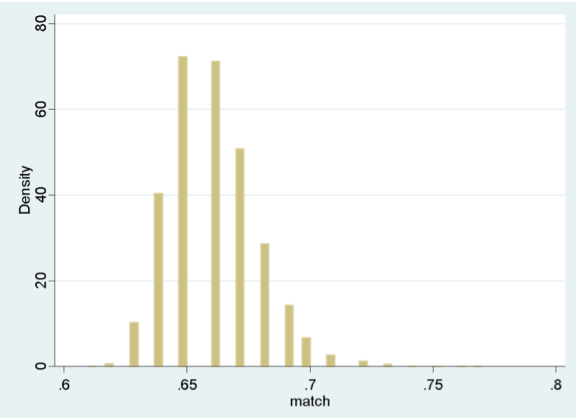
Over 100,000 simulated data sets were created with the following characteristics: 1,000 observations, 100 variables, and random assignment of two distinct values for each cell (0/1). We chose to include only two distinct values in cells to conduct a more rigorous test of the likelihood of duplications since many international surveys use 4- or 5-point scales that should result in lower levels of duplication.

After calculating the maximum percent match between observations and plotting the distributions, we found that the distribution closely resembled a Gumbel curve with a mean of 0.66 ($\beta = 0.15$). Additionally, none of the simulations resulted in a maximum percent match above 85%.

The distribution's mean of 0.66 can be given by the equation $\bar{x} = \frac{2}{a+1}$ where a is the number of options for each variable or question. For example if the simulation were repeated with 3 distinct options instead of 2, the mean value for the maximum match by observation is about 0.5.

Since each response in the simulated data is completely independent and there is no correlation between respondents, simulations cannot closely mirror real-world survey data.

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



However, random simulations can provide a clear picture of what completely independent data would yield.

Additionally, as a robust check on the low likelihood of high percentage matches between observations, we conducted some simulations on correlated data. Simulated data with correlations were constructed with the following correlation structure, represented as a lower triangular matrix in table 1.

Table 1: Simulated Correlation Matrix

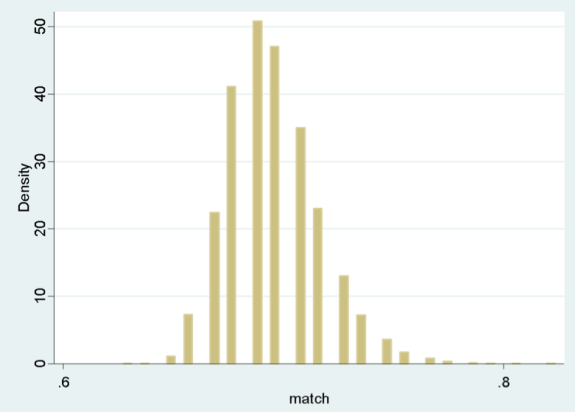
	k_1	k_2	k_3	k_n
k_1	1			
k_2	$C_{k_2k_1}$	1		
k_3	$C_{k_3k_1}$	$C_{k_3k_2}$	1	
k_n	$C_{k_nk_1}$	$C_{k_nk_2}$	$C_{k_nk_3}$	1

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 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 simulates real-world survey conditions, i.e. integer values.

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



After 100 simulations following the steps above, we found that the highest values for percentage match were <85%.

When widely reputable U.S. surveys were analyzed in the same manner as the simulated data, we observed similar distributions for the highest match by observation. For example, the 2014 General Social Survey (after accounting for skip patterns in the questionnaire) and the 2012 American National Election Survey both have probability density functions that mirror the Gumbel distribution in Figure 1. In the two high quality surveys, mean highest percent match was within expected ranges (Social Survey = 0.54; Elections Survey = 0.67) and the highest percentage match value was well below 0.9.

Therefore we can say that the curve given by plotting unfalsified maximum percentage match in survey data should fit the Gumbel curve. The shape of the curve accounts for the fact there is a minimum percentage match between observations; when there are a limited number of possible responses per question (often between 2 and 5), some overlap

between responses is necessary. In the case of simulations with 2 distinct options, the lowest maximum percentage match between observations was 60%. Additionally, the Gumbel curve's long tail on the high end fits with the small likelihood that there is a high resemblance between some observations because of the number of opportunities (usually sample size of 1,000 or more) to match across a number of variables.

The Scope of the Problem

Many organizations do little to check for duplicated responses in surveys. Among those that do, the most common check for exact duplicates across all variables. Checking for exact duplicates among substantive variables is less common, while checking for near duplicates within substantive variables has been performed rarely if at all in the past by major research organizations. Thus, potential falsification via duplication is likely to remain a common problem.

To assess the degree to which this issue persists, we have collected a number of publicly available data sets for international surveys. The surveys come from the Afrobarometer, the Arab Barometer, the Asia Barometer, the Asian Barometer, the European Social Survey, the Eurobarometer, the Latin America Public Opinion Project (LAPOP), Pew Global Attitudes Project, Pew Religion Project, the Sadat Chair at the University of Maryland, and the World Values Survey. A full list of data sets analyzed is listed in Table 2.[4] In total, we analyzed 604 national surveys collected over a period of 35 years.[5]

For each survey, we examined the distribution of maximum percent match for substantive variables for every unique country-year for two primary criteria to determine if

[4] All data sets that were analyzed were downloaded in January 2015.

[5] In the examples that follow we do not list the name of the survey organization nor the country. Instead, we provide some basic background about the country and the survey.

Table 2: Data Sets for Analysis

Survey	**Wave/Year**	**Countries**
Arab Barometer	1	6
Arab Barometer	2	10
Arab Barometer	3	12
Afrobarometer	1	12
Afrobarometer	2	16
Afrobarometer	3	18
Afrobarometer	4	20
Asia Barometer	2003	10
Asian Barometer	2	8
European Social Survey	6	29
Eurobarometer	2001	13
Eurobarometer	2012	29
Pew Global Attitudes	2007	47
Pew Religion Project	2008	19
Pew Religion Project	2012	26
Sadat Chair	2003	6
Sadat Chair	2004	6
Sadat Chair	2005	6
Sadat Chair	2006	6
Sadat Chair	2008	6
LAPOP	2008	24
LAPOP	2010	18
LAPOP	2012	18
World Values Survey	1	10
World Values Survey	2	18
World Values Survey	3	53
World Values Survey	4	41
World Values Survey	5	58
World Values Survey	6	59

there is a significant likelihood of substantial data falsification via duplication.[6] We considered: (1) Is the distribution monotonic on each side of the mode? (2) Are there fewer than 5% of observations where the maximum percent match exceeds 0.9? If both conditions

[6]Items where there were no observations for a given country-year were dropped from analysis.

Table 3: The Scope of the Problem

Degree of Likely Falsification	**Percentage of Surveys**
Low (no cases)	42%
Low (isolated cases)	24%
Moderate	19%
High	16%

are met, then the likelihood of data falsification via duplication is low. If one condition holds but the other does not, then there is a moderate risk of data falsification. If neither condition holds, then our expectation is that the risk of substantial data falsification via duplication is high. Each author coded each country-year independently and we reconciled cases where our coding differed.

Given that survey data sometimes have multiple ballots and there are limited skip patterns in some of the data sets that were analyzed, our coding of the Gumbel distribution sought only to highlight cases where there was a clear violation. As detailed below, distributions that generally approximate a Gumbel, even if there are multiple peaks, were *not* necessarily coded as violating the first condition. Instead, our methodology to approximate the scope of the problem sought only to identify cases where the distribution clearly violated the expected distribution. Our estimates of the data sets affected by duplication are conservative. The results are presented in table 3.

Low-Risk Cases

Examples of distributions that qualify as low-risk are presented in Figures 3 and 4. Figure 3 is a distribution from a recent survey done in an OECD country. The distribution approximates a Gumbel with a modal percent match of slightly less than 0.7. The distribution is monotonic on both sides of the mean and there are only 12 observations greater

than 0.9. Therefore the likelihood of wide-scale data falsification by duplication is small.

However it is noteworthy that there are three exact duplicates present, in addition to nine near duplicates (0.9 <percentmatch <1). Our recommendation would be to remove these observations from the final data set, but given the small number their presence does not overly bias the point estimates or the confidence intervals for statistical analysis. Thus, it is coded as being at lower risk for falsification via duplication.

Figure 3:

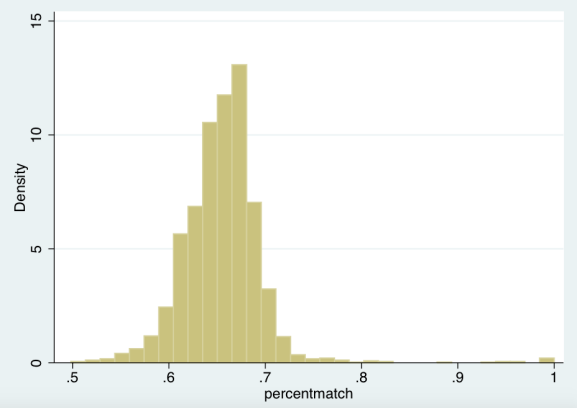
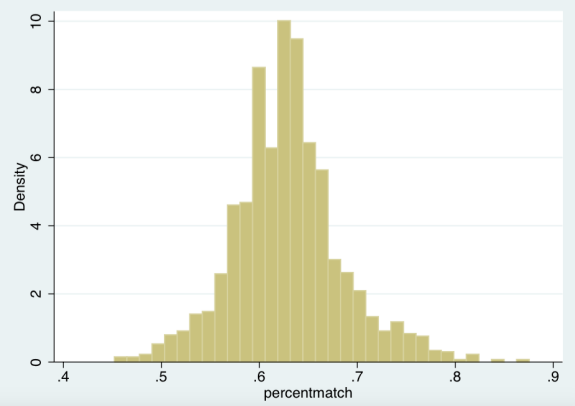


Figure 4 presents another distribution from a study in an OECD country. The distribution resembles a Gumbel with a mode of just over 0.6. The distribution is not strictly monotonic to the left of the mode due to a second peak at 0.6. Nevertheless a Gumbel distribution generally holds on the left side, so this was not coded as a violation of the first condition. Similarly, strictly speaking, the distribution is not monotonic to the right of the mode, with an additional peak around 0.73. However since this minor peak does not significantly depart from the overall trend, again, it is not coded as a violation of the first condition. Since there are no observations with a percent match that exceeds 0.9, we conclude that the likelihood of significant falsification is low for this survey.

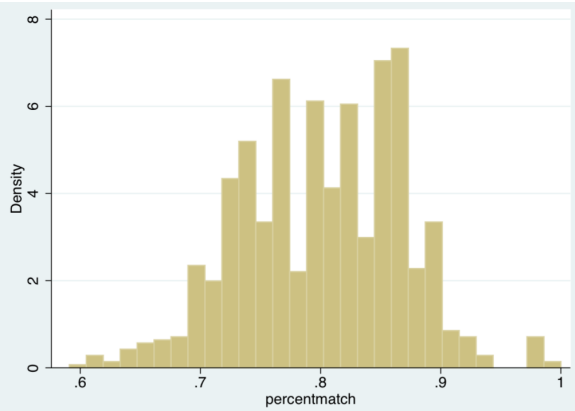
Figure 4:



Medium-Risk Cases

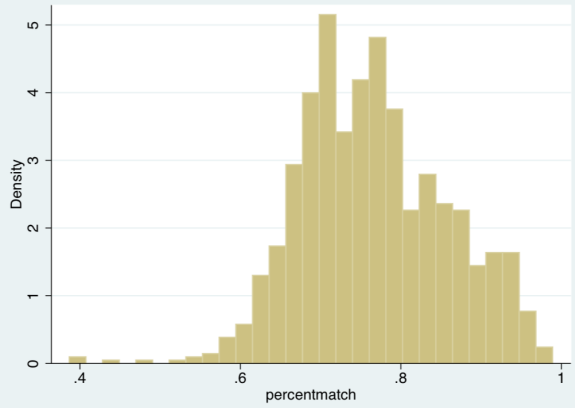
In our examination of the data, there were relatively few cases both where the observation did not resemble a Gumbel distribution but met the second condition. One such case is from a recent survey in an upper-middle-income country, as shown in Figure 5. The distribution does not closely approximate a Gumbel curve; it has seven distinct peaks between 0.7 and 0.9. However, there are also relatively few observations (4%) with a percent match of greater than 0.9. Only one condition is met, suggesting that there is a medium likelihood of falsification via duplication.

Figure 5:



Far more common were cases that more-or-less approximated a Gumbel distribution but had a significant number of observations that exceeded a percent match of 0.9. One such example comes from a study in a lower-middle-income country presented in Figure 6. Although not a perfect Gumbel, this distribution does not have a major peak to the right of the mode. However, it does have a high percentage (10%) of observations with a percent match that exceeds 0.9, as well as a very high percentage between 0.8 and 0.9. Therefore there are reasons to suspect that falsification through duplication has taken place.

Figure 6:

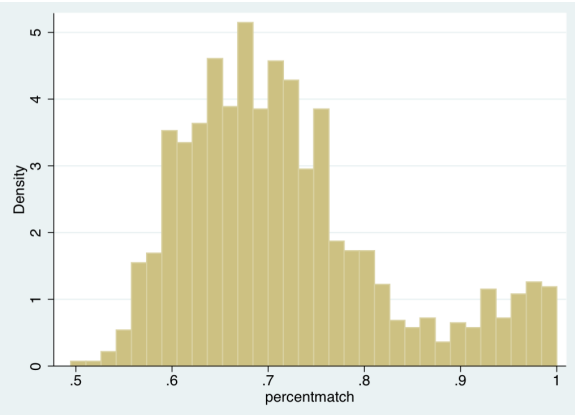


High-Risk Cases

As noted above, a significant percentage of surveys—roughly 16% of those analyzed—reveal a high risk for widespread falsification, exhibiting a distribution that does not approach a Gumbel and a large percentage of observations for which percent match exceeds 0.9. Nevertheless, the degree to which this problem affects surveys varies widely. Figure 7 presents a distribution for a study in a lower-middle-income country with telltale signs of falsification through duplication. In this case the distribution is far from monotonic to the right of the mode, reaching a nadir at just below 0.9 and then increasing once

again. This distribution strongly suggests that either the firm or some of its fieldworkers deliberately engaged in falsification via duplication. For the substantive variables, only three observations are exact matches. However, it strongly appears that at least 10% of observations in the survey were deliberately modified in an attempt to conceal falsified data.

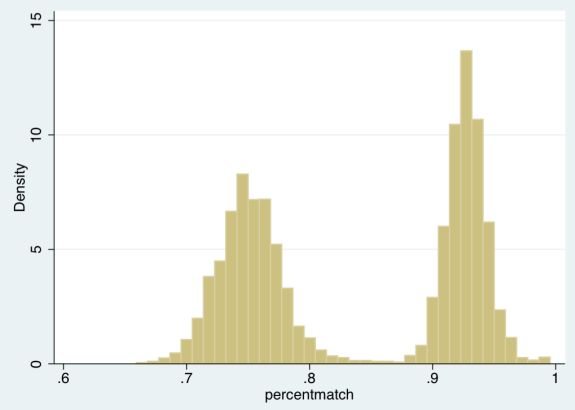
Figure 7:



A more brazen attempt at falsification is observed in a survey from an upper-middle-income country presented in Figure 8. In this example, the distribution is bimodal with the first peak at around 0.75 and the second at roughly 0.92. Based on this distribution, it is very probable that the local firm carried out valid interviews for nearly half of the observations, but increased the sample size via duplication. In total, half of the observations have a maximum percent match in excess of 0.9. The firm appears to have been careful, however, to ensure that there were no exact duplicates.

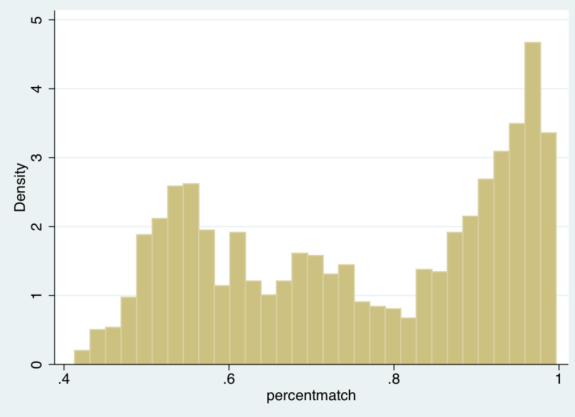
Another example of a survey with a high probability of widespread falsification comes from a lower-middle-income country as seen in Figure 9. In this case, the distribution does not approximate a Gumbel and a third of the observations have a maximum percent match greater than 0.9. However, it should also be noted that the problem with falsification does

Figure 8:



not appear to begin at the 0.9 threshold; instead, the distribution begins to increase at roughly 0.8 (48% of all observations) and peaks at slightly less than 1. Moreover, even below 0.8 the distribution does not approximate a Gumbel, potentially calling into question the validity of interviews below this threshold. Thus, it is very likely that this survey suffers from widespread falsification via duplication.

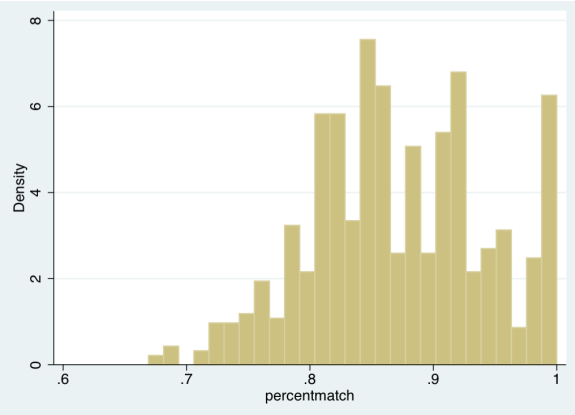
Figure 9:



Another study in Figure 10 of a lower-middle-income country reveals a distribution that suggests widespread falsification. Again, the distribution does not resemble a Gumbel

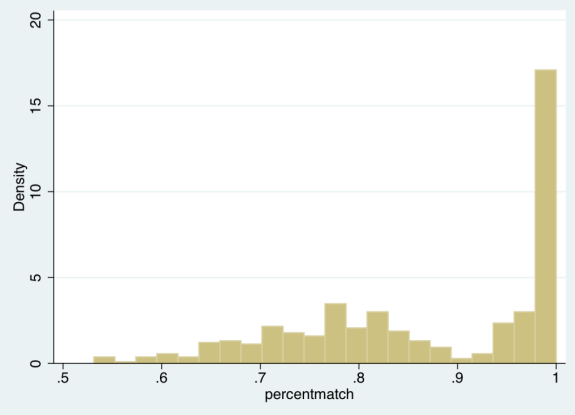
distribution and 36% of observations have a percent match that exceeds 0.9. Additionally, 85% of observations exceed a percent match of 0.8, implying massive falsification.

Figure 10:



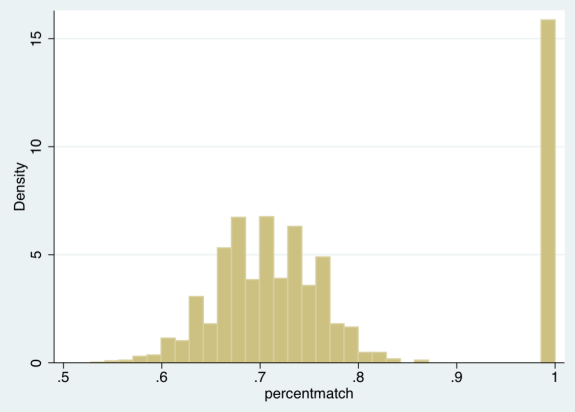
Based on the data sets we examined, firms do not always attempt to conceal falsification through the use of near duplicates. Figure 11 presents the distribution from a study in a high-income non-OECD country. The distribution does not approximate a Gumbel due to 26% of all observations being exact duplicates on substantive variables. Moreover, fully half of observations have a maximum percent match exceeding 0.9.

Figure 11:



Notably, the problem of falsification via duplication is not limited to countries where survey research is a new field or where there are few survey firms. Figure 12 presents a study from an OECD country where 23% of observations are exact duplicates for substantive variables. Excepting these exact duplicates, the remaining observations somewhat approximate a Gumbel, and there are no other observations for which the maximum percent match exceeds 0.9. Thus, it appears that real data underlie the majority of observations, but the high percentage of exact matches is likely to significantly bias point estimates and confidence intervals, meaning it is essential they be dropped from any analysis.

Figure 12:



Using *percentmatch* to Detect Falsification[7]

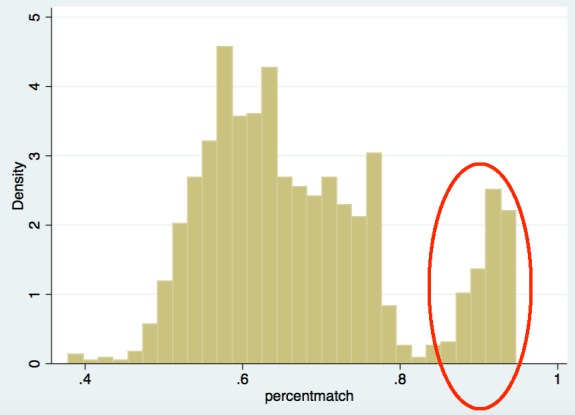
As the previous section demonstrates, falsification through duplication remains a common strategy of unscrupulous firms and interviewers, but we offer *percentmatch* as a tool that researchers can use to quickly and effectively estimate the degree to which duplication impacts a survey project. In effect, it raises the costs and challenges of falsification, meaning firms are less likely to falsify in this manner and more likely to ensure that there

[7] *Percentmatch*, can be downloaded from SSC (the Statistical Software Components archive) or from SSRN. Cite use of the program or our method by referencing this paper.

is sufficient oversight of their field force. Of course, although it makes falsification more difficult, this does not mean an end to fraudulent practices.

Beyond detecting falsification, *percentmatch* represents a valuable tool for detecting why falsification has taken place. For example, the Arab Barometer team has used this tool to great effect to improve our data quality control process. We recently fielded a survey with a local partner and used *percentmatch* on the raw data file. The original distribution is presented in figure 13.

Figure 13:



The results show telltale signs of data falsification via duplication when the maximum percent match exceeds 0.8. The team immediately began to more closely examine these observations to determine what went wrong. Comparing these results by interviewer yielded a clear common link between the observations as presented in table 4.

The results reveal a strong correlation between the high percent match observations and interviewer 1, with all but one of his 123 interviews exceeding a percent match of 0.8. For four others, a minority of interviews also exceeded 0.8. Closer analysis revealed that interviews conducted by interviewer 1 were linked with those of other interviewers, suggesting that this person used potentially valid interviews conducted by others to falsify

Table 4:

Interviewer	*Match* $< 80\%$	*Match* $\geq 80\%$
1	1	122
2	127	0
3	72	0
4	63	0
5	113	0
6	77	28
7	115	0
8	76	0
9	37	0
10	34	0
11	61	4
12	3	0
13	84	0
14	58	0
15	62	11
16	39	13

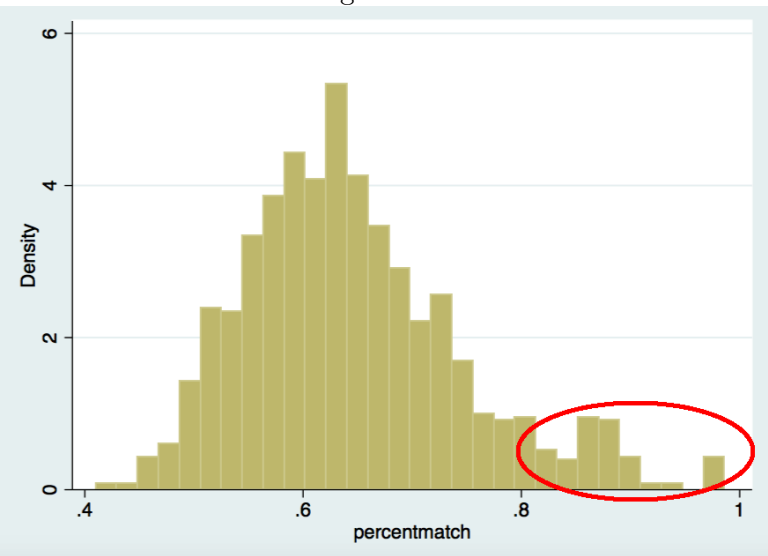
his interviews.

It is notable that no observations collected by interviewer 1 had a maximum percent match exceeding 0.95, meaning that this problem would likely have gone undetected if standard checks were applied. Further analysis revealed that interviewer 1 attempted to conceal the falsified interviews by randomizing the responses for roughly every 10^{th} variable. Thus, his interviews were unlikely to produce long strings of duplicates that might be detected through looking for exact matches for a string of consecutive variables.

By all indications the data for the remaining interviewers appeared to be validly collected. Expected correlations held between variables in the data set and, although not a perfect Gumbel, the distribution is not one that would lead to the expectation of widespread falsification for the remaining observations. Thus, all observations with a percent match exceeding 0.8 were eliminated from the data set prior to publication.

A second example yields a more challenging form of falsification, which the Arab Barometer team was able to detect by using *percentmatch*. This study yielded the distribution seen in figure 14.

Figure 14:



This distribution presents a significantly lower expectation for data falsification than the distribution in figure 13. In this case, only 1.5% of all observations have a percent match that exceeds 0.9. Moreover, the general distribution approximates a Gumbel. Yet, on the right tail there are small but discernible peaks when percent match exceeds 0.8. Thus, the Arab Barometer team examined these observations with greater scrutiny.

Table 5 presents instances when percent match exceeds 0.8 by interviewer.[8] 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 0.8.

Although we examined all of these interviews, the Arab Barometer team focused especially on those done by interviewers 8 and 9. The results showed that for the vast

[8]The table includes only interviewers with observations with a percent match $\geq$ 80%.

Table 5:

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

majority of variables there were no discernible differences between the observations of these two interviewers and all others. However, this pattern did not hold for variables with a religious dimension. The data reveal that respondents interviewed by interviewers 8 and 9 were far more religious and supportive of political Islam than those of all other interviewers. Table 6 shows support for the main Islamist party in the country for interviewers 8 and 9 versus all other interviewers.

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

Interviewers 8 and 9 did not conduct interviews in areas where Islamist party support is overly strong, whether based on election results or informed opinions from local scholars. Additionally, the Arab Barometer team was able to identify a second data set where these two interviewers conducted interviews on a similar range of topics. This second data set revealed a similar pattern where support for political Islam was particularly strong in observations conducted by these two interviewers.

Based on this analysis, the Arab Barometer team concluded that these two interviewers either 1) falsified the data for variables relating to religion and political Islam due to their normative biases or 2) selected respondents through a non-random process that would yield this outcome. In either case, these observations do not represent randomly-selected valid observations and were removed from the data set prior to publication.

Potential Limitations

There are potential limitations to this program that users should take into account. First, items that were not administered to all respondents should be removed from the data set prior to using *percentmatch*. Not accounting for skip patterns or split ballot questionnaires will result in an artificially high share of near duplicates. Analysis should either be conducted separately for each half of the split sample or should be run excluding variables unique to the portion of the sample dropped from the data set.

Second, *percentmatch* is more effective with a larger-n survey and a lengthy survey instrument. The greater the number of both observations and effective variables, the likelier it is that the distribution will approximate a Gumbel.

Finally, this approach to identifying fraudulent observations in surveys does not easily translate outside the world of international social science surveys. For customer satisfaction or employee engagement surveys, where every question is 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. Part of the reason near duplicates are highly unlikely in social science public opinion polling is the diversity of questions and topics.

Conclusion

Our summary table of the degree to which falsification is present in existing publicly available data sets would have considered the distribution in figure 14 to be a relatively low risk of falsification. Yet, under greater scrutiny, it became clear that this data set also suffered from a limited degree of data falsification. This example makes clear the opportunity for researchers to use *percentmatch* as a tool to more rigorously examine survey data for instances of potential falsification.

Often, survey researchers want to know what percent match is too high to be considered valid. Although data sets with a high percentage of near duplicates (controlling for skip patterns) strongly suggest falsification, it is important to recognize that focusing on the histogram actually reveals more about the overall quality of the data. Researchers should examine surveys where the distribution departs from a Gumbel distribution more closely in order to evaluate the overall quality.

While it may be possible to examine the overall distribution and determine whether falsification is present, understanding the reasons for falsification is critical to limiting its likelihood in the future. This process takes time and exploration of the data. It is not always clear why falsification takes place. As the experience from the Arab Barometer indicates, it may not always be the result of shirking but may also be introduced for other reasons, such as the normative biases of interviewers.

With sufficient analysis, researchers will often be able to tell if falsification was introduced by individual interviewers while in the field or by the firm while compiling data. If duplication is distributed across the work of many individual interviewers, then it is reasonable to infer the firm bears a significant degree of culpability, and survey researchers should avoid doing business with them in the future. In the case that falsification is more limited and linked to specific interviewers, researchers can work with the firm to improve

their training and oversight procedures.

Indeed, the consequences of data falsification are significant. Near duplicates with a higher number of cases provide a large base sample size for statistical significance testing and solidify the relationships between variables, all while reducing variance. Ultimately, this means more stars for model coefficients in journal articles—boosting signal and taking away noise.

Although we encourage survey researchers to look at their past surveys to ensure data meet the highest quality standards, *percentmatch* 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 better estimates of public opinion around the world.

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 data sets that have not been scrutinized for quality by programs like ours may not not hold up after accounting for fraud or error. 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) or from SSRN. 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.

Cohen, William W. 1998. "Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity." *Proceedings 1998 ACM SIGMOD International Conference on Management of Data* pp. 201–212.

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

Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis and Vassilios S. Verykios. 2007. "Duplicate Record Detection: A Survey." *IEEE Transactions on Knowledge and Data Engineering* 19(1):1–16.

Jaro, Matthew A. 1976. *Unimatch: A Record Linkage System: User's Manual*. U.S. Dept. of Commerce, Social and Economic Statistics Administration, Bureau of the Census.

Monge, Alvaro E. and Charles P. Elkan. 1996. "The Field Matching Problem: Algorithms and Applications." *Proceedings Second International Conference on Knowledge Discovery and Data Mining* pp. 267–270.

Mushtaq, Ali. 2014. Detection Techniques Applied. In *Conference presentation at Washington Statistical Society's "Curb-Stoning: a too neglected and very embarrassing survey problem"*.

Navarro, Gonzalo. 2001. "A Guided Tour to Approximate String Matching." *ACM Computing Surveys* 33(1):31–88.

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

Smith, T.F. and M.S. Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147:195–197.

Taft, Robert L. 1970. "Name Search Techniques." *Technical Report Special Report No. 1, New York State Identification and Intelligence System* .

Ukkonen, Esko. 1992. "Approximate String Matching with q-Grams and Maximal Matches." *Theoretical Computer Science* 92(1):191–211.

Ullmann, J.R. 1977. "A Binary n-Gram Technique for Automatic Correction of Substitution, Deletion, Insertion, and Reversal Errors in Words." *The Computer Journal* 20(2):141–147.

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.

Waterman, M.S., T.F. Smith and W.A. Beyer. 1976. "Some Biological Sequence Metrics." *Advances in Mathematics* 20(4):367–387.